In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institute shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.

7/25/68

# USE OF MULTIPLE DISCRIMINANT ANALYSIS TO EVALUATE THE EFFECTS OF LAND USE CHANGE ON THE SIMULATED YIELD OF A WATERSHED

A THESIS

Presented to

The Faculty of the Graduate Division

by

Donn Gene DeCoursey

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in the School of Civil Engineering

Georgia Institute of Technology

June 1970

USE OF MULTIPLE DISCRIMINANT ANALYSIS TO EVALUATE THE EFFECTS OF LAND USE CHANGE ON THE SIMULATED YIELD OF A WATERSHED

Approved:

Chairman

Date approved by Chairman: May 28, 1970

# ACKNOWLEDGMENTS

This work was performed under the supervision of Professor Willard M. Snyder. It was through consultation with him that the research project was conceived. The author also acknowledges the many valuable comments and suggestions that he made.

The author wishes to acknowledge the assistance of the Agricultural Research Service and the Director of the Southern Plains Watershed Research Center, Mr. Monroe Hartman, for providing the computer assistance, facilities, and time needed to complete the project. The Director of the Blacklands Experimental Watershed, Mr. Ralph Baird, is also to be acknowledged for his assistance in providing the data upon which this research project was based.

The author would also like to acknowledge the assistance of Mr. Edward Seely who wrote and checked many of the computer programs needed for the research. His many suggestions were also appreciated. The able assistance of Mrs. Betty Golden, who typed the thesis was very much appreciated.

The author would, in particular, like to thank Dr. T. W. Jackson, who at that time was the Acting Dean of the Graduate Division of the School for granting permission to complete the research off campus.

Lastly, I would like to thank my wife Shirley who gave me the unyielding support needed to complete the work.

TABLE OF CONTENTS

# LIST OF TABLES

Page

Table

Page

Table

Page

Table

Page

Table

LIST OF FIGURES

Page

Figure

## SUMMARY

Multivariate statistical techniques have been used successfully in hydrologic analyses, however, no emphasis has been placed on the use of multiple discriminant analysis. It can be used to advantage in evaluating differences in hydrologic data because sets of data may be analyzed as a unit rather than individually.

The hydrologic problem selected to demonstrate the technique is that of evaluating the hydrologic effects of a change in the land use parameters of a watershed model and determining the significance of length of hydrologic record.

Synthetic records, 2, 10, and 30 years in length, were used in the analysis. Synthetic data were generated because the effect of climatic variability could be eliminated to different degrees by studying several sets of records of different length.

A watershed, 1,110 acres in size, located in the Blacklands of Texas, was selected for fitting to a watershed model. The watershed model was a hyperbolic functional relation between rainfall and runoff combined with a threshold concept. Runoff as defined by the model is a function of storm rainfall, pan evaporation, soil moisture, and land use. The initial abstraction is a function of the antecedent soil moisture and land use. A bookkeeping technique is used to get a continuous record of soil moisture. Parameters of the model have been determined for five different land uses; native grass meadow, Bermuda pasture, cultivated row crops, cultivated-oats, and cultivated-no crops.

A probabilistic element equal to the unexplained variance of the model was added to each calculated value. The statistical characteristics of the probabilistic element are functions of the size of the calculated runoff event.

A technique for generating synthetic daily rainfall and pan evaporation data for input to the model was developed. The temporal distribution of rainfall was calculated by a two-state Markov chain of transition probabilities. Rainfall amounts were found to be independent and described by a skewed log normal distribution. Evaporation data were generated by using a linear equation which was a function of rainfall and previous evaporation.

Linear discriminant functions which were used to analyze the hydrologic differences between land use groups are defined such as to produce the largest possible difference between the groups. It is established by maximizing the ratio of the between-groups sums of squares to the within-groups sums of squares. The number of functions is either the number of predictors or one less than the number of groups whichever is less. A classification technique is used to assign the individual observations to a group based on the discriminant scores.

Results of using the discriminant analysis on the different land use groups show that it can be used to distinguish between groups and isolate those variables most indicative of group differences. When applied to the results of the model, the discriminant analyses indicated that the significant discriminators were characteristics of runoff from small rainfall events. They also indicate that extensive changes in land use parameters will probably be distinguishable in summaries of

moderate to long periods of record. In general it is also evident that the percent of a watershed involved in land use change is more important than the percent of watershed in specific land uses. Tests of the scheme used in generating the synthetic input data for the model show that it produced data with the same statistical characteristics as the historic data.

# CHAPTER I

# INTRODUCTION

## Problem Statement

The problem to which this dissertation is addressing itself is that of introducing a statistical tool for use in the analysis of hydrologic data. Many problems of a hydrologic nature could be selected to demonstrate the applicability of the technique. However, the significance of land use change as related to watershed yield is of particular interest to many hydrologists, especially those who are having to rely on watershed models to estimate the hydrologic characteristics of the watershed under consideration. Therefore, the problem of determining the hydrologic significance of land use change as indicated by a watershed model was selected to demonstrate the statistical technique.

The statistical method, multiple discriminant analysis, a technique of multivariate analysis, has been used primarily in the fields of biometrics and psychology. In these fields, individuals, objects, or phenomena have been described by batteries of tests or measurements and the results substituted into a series of linear functions. The values of these functions, discriminant scores, have then been used to classify the objects into discrete groups. Using this approach, an object may be assigned to the group that the characteristics of the object indicate it is most nearly like. No other statistical technique yet developed will do precisely this.

## Justification

The availability of a supply of potable water is of increasing importance as evidenced by the activity of the Federal and State Governments as well as many other organizations. One of the many variables affecting this supply of water is vegetation growing on the land. Changes in this vegetal cover (land use) produce different effects on the hydrology of a watershed depending upon other characteristics of the watershed and the climate of the area. The magnitude of hydrologic change is a controversial question. It is known that in humid climates changes in land use, i.e., changing from pasture to cultivated crops, do not materially affect yield because evapotranspiration is at or near potential rate most of the time and plant growth is not limited by the amount of moisture present in the soil. However, some changes in land use such as clear cutting of timber can make a significant change in the hydrology.

In the subhumid climates, moisture for plant growth is an almost ever-present problem. Therefore, if land use changes require additional moisture, it will be used by the plants and is not available for other uses. Under these conditions land use changes are more significant than in humid areas.

In March of 1957 a study was initiated by the Bureau of Reclamation, Soil Conservation Service, and Agricultural Research Service, to determine the magnitude of conservation activity on the yield of a watershed. The study was primarily concerned with the dry subhumid-to-arid areas such as the Great Plains, Midwest and Southwest parts of the United States. Following are two excerpts from the final

report of the study (1):

> Many methods of evaluating effects of watershed treatment on streamflow were tried. Included were simple correlations and regressions, multiple correlations and regressions (linear and curvilinear), analyses of variance, time-series studies, double-mass diagrams, "before and after" comparisons, hydrograph analyses, and others. All the methods were basically directed toward determining if there had been changes in the precipitation-streamflow relations of the river basins. No statistical approach was found that would consistently assess effects of land treatment on streamflow from river basins, or even prove conclusively that such effects do or do not exist. In a few cases, streamflow appeared to be increasing. In some, it appeared to be decreasing. In all cases, streamflow fluctuated considerably, due to climatic or other causes. This lack of positive findings should not be interpreted to mean, however, that the conservation use and treatment of upstream land has no effects on downstream water yields by streamflow. It is axiomatic that there must be such effects in dry subhumid-to-arid areas where available soil moisture, not solar energy, consistently limits evapotranspiration.
>
> Overall, the many investigations carried out demonstrated that a procedure, based only on statistically significant results obtained from studies of river basins and research watersheds, could not be developed. Yet, the evidence on the whole indicated that conservation measures such as contouring did affect on-site runoff, and it is self-evident that, in drier areas, storage in ponds and reservoirs, drainage of potholes, and irrigation affect on-site water yield. Since a procedure could not be demonstrated statistically, a rational procedure was developed. The best available information relating to the various components of the hydrologic process involved in the generation of precipitation excess and delivery of water yields by streamflow was the basis for the procedure.

Even though a procedure was developed to evaluate the effect that land use and conservation practices had on yield, they were not able to show statistically that there were significant effects on the yield of a river basin. However, the authors may have recognized one of the reasons for the lack of statistical significance, "It may also be that the statistical models used to analyze these data were not appropriate. Indeed, it is strongly suspected that the seemingly ideal statistical model - multiple regression - is not applicable to the

hydrologic data now available." They also noted one other item which could have contributed to the lack of statistical significance:

> Furthermore, most of the conservation work has been accomplished in the last few years, and thus has had a very short time in which to function. A very large change in the precipitation-streamflow relation in a few late years in a long streamflow record will not show up statistically significant, so great are the variances involved.

These statements would indicate that most hydrologic records are not long enough to show statistically significant changes in precipitation-streamflow relations caused by changes in land use primarily because climatic variation exerts a much stronger influence on the relation than does land use change. To get around this problem, one must resort to some method of eliminating the effects of climatic variation. The method selected for this study is based on using available data and an analytical model. The analytical model is composed of two parts; the first being a scheme for generating synthetic input data, and the second being a watershed model. The problem of evaluating the significance of land use change thus degenerates from one of evaluating land use changes directly from a watershed to one of evaluating it as it is defined by the model selected.

However, the use of a watershed model to evaluate land use change has one distinct advantage over a direct evaluation; that is that most of the data used in project design is synthetic; i.e., based on the output from some analytical model. It would therefore be of benefit to know whether or not watershed models should be a function of land use considering the fact that any model, no matter how complex, explains only a part of the total variation of the observed record.

The statements by Sharp et al (1) also indicate that in the past a change in the hydrology of a watershed has been investigated on the basis of one hydrologic element at a time. However, hydrologic change may be a more subtle function of several variables. If this is true, then detection and measurement of change must be approached on the basis of a before and after comparison of sets of variables, not one. In these sets the variables may not be statistically independent and normally distributed. Therefore, many standard statistical tests of significance are inappropriate.

Recent successful applications of multivariate statistics in hydrology have been made. However, no emphasis has been placed on the use of multiple discriminant analysis. It appears to be ideal for the purpose of analyzing sets of non-independent variables used for group distinction such as the detection of hydrologic change.

## Objectives of the Investigation

The objectives of this dissertation are, in order of importance:

(1) Present and demonstrate the use of multiple discriminant analysis in the study of hydrologic data.

(2) Determine the effect that the degree of land use change has on the ability to distinguish hydrologic differences in a modeled watershed.

(3) Determine the effect that length of record has on the ability to distinguish hydrologic differences in a modeled watershed.

(4) Develop a technique for generating synthetic climatic data.

The first objective is in part a response to the problem

introduced by Sharp, et al (1) in the statement that multiple regression may not be applicable to some types of hydrologic data analysis. The variables affecting the yield of a watershed are highly interrelated and as such have a high degree of correlation among themselves. One of the premises upon which multiple regression analysis is based is an assumption of independent variables. If the variables are not independent, then the parameters of a multiple regression analysis, even though they are "good" in a statistical sense, cannot be evaluated objectively. This is part of the problem that Sharp faced. Multivariate analysis is a branch of statistics in which the orthogonal or independent components of a system are isolated and used in the analysis of the system. One technique is presented and used to distinguish differences in watershed hydrology under different land uses.

The second and third objectives are an attempt to answer the general question, "Does land use change produce a statistically significant change in watershed hydrology when the watershed hydrology is defined by an analytical model?" This research can, however, answer the question only when posed in a very restricted sense, i.e., with reference to the land use, climate, soils, geology, and size characterized by the watershed selected for study. The results are also only applicable for watershed models similar to the one selected for use in the study. Other considerations involved in meeting the second and third objectives are:

(1) On very small watersheds, those in one crop for example, land use and treatment can reduce surface runoff from 25 to 40 percent particularly in dry years, but on large watersheds it is very hard and

at time impossible to find such effects.

(2) Changes in land use are qualitative rather than quantitative. As such, the degree of change is subjective as well as relative, and impossible to define in quantitative terms.

(3) Identical land use changes can have different results depending upon the geographic location, climate, soils, and other characteristics of the watershed and whether or not the record was taken from a wet or dry period. The following quotation from Sharp et al (1) brings this point into focus:

> Applying logic to the water-yield problem indicates that water yields are residuals from precipitation after the demands of evapotranspiration are met. In humid to perhumid areas, evapotranspiration is near potential evapotranspiration as limited only by the solar energy available; the vegetation seldom suffers from protracted periods of soil moisture stress, because frequent and adequate precipitation keeps soil-moisture quantities at relatively high levels.
>
> In arid areas, on the other hand, available soil moisture, and not solar energy, limits evapotranspiration. Vegetation suffers nearly every year, and for protracted periods, from high soil-moisture stress. In dry subhumid areas, most years will be dry enough that lack of soil moisture limits evapotranspiration. In moist subhumid areas, most years will have largely adequate soil moisture, and it is only in the drier years that evapotranspiration will be markedly limited by soil-moisture exhaustion.
>
> From the above and from known effects of the conservation use and treatment of land on on-site runoff, it is apparent that the conservation use and treatment of land, as we know it today, will have only very limited effects on water yield from large watersheds in humid climates. In dry subhumid-to-arid climates, there will be effects from these practices. In moist subhumid climates, there will be no effects in the wet years, but there may be effects in dry years.

(4) As was mentioned briefly in the preceding discussion, the cost of many construction projects is based directly on some hydrologic characteristic. It may be the quantity or peak rate of flow from some

watershed. Much of the time no information is available on the actual flows from the watershed in question, and they need to be estimated by some form of watershed model. The Soil Conservation Service project formulation computer program is probably used for estimating stream flow more than any other method. One of the watershed characteristics affecting this model or any other one selected is the land use. The previous discussion indicated that the significance of land use change is not known.

If the significance of land use change cannot be determined in the real world situation, then is it significant when included as part of a watershed model? This question is especially significant when one considers that no model completely describes the hydrology of a watershed. There are always unexplained discrepancies between the model and the real world situation. If the model is very poor, then the unexplained variation could completely mask any change in the model output caused by a change in land use.

The fourth objective is primarily a by-product of the system as it was needed to generate input for the watershed model.

## Approach to be Used in the Investigation

The following approach was used in meeting the objectives of the investigation:

A watershed model which would predict storm runoff was fitted to a watershed with a minimum of 20 years of stream flow records encompassing at least two different land use conditions.

Characteristics of the rainfall and other climatic data used in

the model were obtained from data on the watershed under investigation. A system which would generate data with these same statistical properties was then developed. The generation scheme was combined with the watershed model to calculate synthetic sequences of storm runoff.

Land use characteristics of the watershed were then fixed and several 30-year sequences of runoff generated. Various characteristics of the artificial sequences such as average runoff, percent runoff, etc. were collected. The land use parameters were then changed and several more sequences and their characteristics generated. This pattern was repeated until seven land use patterns had been used.

The characteristics of these seven groups were then subjected to components analysis to find the characteristics most responsive to changes in land use. The most significant characteristics were then used in a multiple discriminant analysis computer program to calculate the discriminant functions maximizing the ratio of the among to the within groups sums of squares. The discriminant scores from this computer program were used along with the group means and variances in a classification program to classify the individual observations into the group to which they were most nearly like. By knowing the group in which the observation belonged, the accuracy of the calssification could be determined. The number of correct classifications compared to the number of wrong classifications was an index of the ability to discriminate between the different land uses.

# CHAPTER II

## LITERATURE SURVEY

In the following discussion mathematical formulae and statistical tests are omitted because in most cases they are quite involved and inclusion would destroy the trend of thought. All of the tests and the mathematical treatments needed in the study are described in the body of the thesis.

### Multiple Discriminant Analysis

The technique of multiple discriminant analysis was originally developed by Fisher (2) in 1936. Since that time, several good texts, Rao (3), Kendall (4), and Anderson (5), have been written on multivariate statistical analyses with chapters devoted to discriminant analysis. The computational requirements of these analyses have resulted in the development of digital computer programs. An excellent description of such programs is presented in Cooley and Lohnes (6) and in a technical report by Cosetti (7). Fisher's original work on two groups is extended to multiple groups in two papers by Bryan (8,9). An excellent mathematical treatment of multiple discriminant analysis is presented in a paper by Miller (10). Extensive bibliographies on the use of the technique are presented in two papers; Hodges (11), and Tatsuoka and Tiedeman (12). The list of "Other References" includes a fairly recent bibliography on use of the technique. Both Hodges (11) and Tatsuoka and Tiedeman (12) also have good reviews of the historical development.

They were used extensively in the following review.

Discriminant Functions

Assuming that several groups of data are available; i.e., runoff characteristics such as peak rate, volume, percent runoff, duration of runoff, etc. on a given number of storms from several different watersheds, it would be desirable to know: (1) Are there differences between these watersheds, (2) if so, how great are the differences, (3) what characteristics distinguish these differences, and (4) given the aerial response characteristics of a storm, could it be properly assigned to one of the watersheds? Multiple discriminant analysis and associated tests will help answer these questions.

Two-Group Discriminant Functions. Fisher (2) defined what he called the discriminant function as a linear combination of variables which would, better than any other linear combination, discriminate between two groups. In so defining this combination, he maximized the ratio of the among-groups sum-of-squares to the within-group sum-of-squares. He showed that the difference between the group means in the discriminant space is proportional to Hotelling's (13) $T^2$ which is a generalization of Student's t-statistic to multivariate cases, and is also proportional to Mahalanobis' (14) $D^2$ which is a measure of the "distance" between the two groups. Hotelling's $T^2$ is used to answer the question "Is there a difference between group centroids in multivariate space?" Welch (15) in 1936 showed that Fisher's discriminant function was the optimum solution to the problem of maximizing the probability of correctly classifying observations into groups, thus paving the way for the extension of discriminant analysis into the

field of phychological testing.

Wallace and Travers (16) made the first application of the method in 1938 in a study of specialty salesmen. Many others such as Selover (17) in studying test scores, Kuder (18) in accounting, Baten and Hatcher (19) in analyzing student ability, and Harper (20) in the classification of schizophrenic groups also used the new technique.

Due to the similarity between discriminant functions and biserial or point biserial regression functions, Wherry (21) advocated the abandonment of discriminant functions because of computational difficulty. Bryan (9), Rulon (22,23), and Tiedeman (24), however, pointed out that the proportionality between the two approaches holds only when comparing two groups at a time.

Multi-Group Discriminant Functions. It is not clear who first extended Fisher's work to include more than two groups. It was used in its multi-group form by several researchers such as Barnard (25), Day and Sandomire (26), Fisher (27), Johnson (28), and Mather (29). However, Bryan (8,9) presented the first workable computational routine for getting the roots and vectors of the matrix involved. Bryan also pointed out that the first root and vector maximize the discriminant criterion in Fisher's original sense; the second root and vector maximize the ratio of the residual among-groups sum-of-squares to the residual within-group sum-of-squares; and so on with the remaining roots. The successive linear combinations of the variables, the coefficients of which are the eigenvectors of the matrix, are called multiple discriminant functions. The number of functions will always be either the number of variables or one less than the number of groups,

whichever is smaller. Both Hodges (11) and Tatsuoka and Tiedeman (12) have extensive bibliographies on the use of multiple discriminant analysis.

Multivariate Tests of Significance

The statistical tests of significance available for application in multiple discriminant analysis assume the group density functions to be multivariate normal with equal dispersions. A statistic used to test this assumption and other multivariate tests of significance are described below.

Hotelling's $T^2$. Hotelling's (13) $T^2$ statistic is a function of the within-groups sum-of-squares matrix and the vector of variable deviations from their means. It was developed primarily to test the difference between group means. The test statistic is distributed as F.

Wilks' Lambda. One year after Hotelling developed the $T^2$ test statistic, Wilks (30) solved its multi-group extension. The test statistic, lambda, $\Lambda$, is the ratio of the determinant of the within-groups sum-of-squares matrix to the determinant of the total sum-of-squares matrix. Bartlett (31,32) found that a function of $\Lambda$ was distributed as $\chi^2$, whereas, Rao (3) transformed $\Lambda$ such that it was distributed as F. Rao's test is the algebraic equivalent of the familiar univariate F test for the one-variate case. Lohnes (33) found by Monte Carlo methods that the latter test was slightly better.

Mahalanobis' $D^2$. Mahalanobis (14) developed the test statistic $D^2$ in 1927. It is directly proportional to Hotelling's $T^2$, but is a measure of distance between groups rather than a criterion for testing the hypothesis of zero-distance as is $T^2$, Rao (3) extended Mahalanobis'

$D^2$ to situations of more than two groups, thus it can be used to test the significance of predictors on an overall group separation basis.

Testing Significance in Multiple Discriminant Analysis. The significance of group difference based on a vector of variable means, is often encountered in research. Hotelling's (13) $T^2$ was specifically designed for this test. Wilks' (3) $\Lambda$, which is the multi-group extension of the $T^2$ statistic, is applicable to any number of groups in testing overall group difference. Rulon and Brooks (34) have summarized several other equivalent ways of testing the significance of differences for the two-group case.

The $\Lambda$ test of the null hypothesis of the equality of mean vectors assumes that the group dispersions are equal and from multivariate normal populations. Bartlett (35) developed a test for the null hypothesis of equality of group dispersions based upon the determinants of the dispersion matrices. Box (36) presented the test and showed that the test statistic was distributed as F. The test is quite involved as is that of $\Lambda$.

In discriminant analysis the ability to discriminate between groups can generally be improved by the addition of more predictors. There is a point, however, where the improvement in discriminating ability is not significant. Rao (3) has shown that the difference in Mahalanobis' $D^2$ caused by the addition of new predictors is distributed as $\chi^2$ and thus can be used to test the significance of the additional predictors. Miller (10), using the $D^2$ statistic as a criterion, developed a method of stepwise selection of predictors analogous to the stepwise selection of predictors in multiple regression. He used Rao's

test statistic as the indicator of statistical significance.

Wallis (37) in studying two groups compared five different methods of selecting predictors: (1) All variables, (2) stepwise selection, (3) unrelated measurement, (4) reduced rank, and (5) factor score method. The last three are based on a principal components analysis and varimax rotation of the factor weight matrix using a dummy variable to distinguish group membership. He found that the unrelated measurement and reduced rank selections were superior to the others especially when tested on a control set of data not used in getting the linear discriminant function.

The number of discriminant functions that are statistically significant is a function of the size of the roots. Miller (10) states that no exact test exists for judging the statistical significance of the roots associated with the individual discriminant functions. Rao (3), however, presents two approximate procedures, one due to Bartlett, for testing the significance of the roots. The test statistics for both procedures are functions of the roots and are distributed as $\chi^2$.

Classification Procedures

Classification techniques are used in assigning an observation to one of several groups in order that group size and membership may be analyzed. Hossock (38) reviewed seven of the more common classification techniques presenting in his review the method of selecting variables, the selection of an estimation procedure, the determination of the classification rule, the measure of effectiveness, the application to an unknown observation, and an example of its use. In this thesis, classification and assignments are based upon discriminant

scores and a set of hypotheses regarding group membership. Classification of observations based upon these criteria were first used in the fields of toxonomy, career guidance, and the selection and assignment of manpower. Working on profile comparison problems, duMas (39) and Cattell (40) developed coefficients of profile similarity, while Cronbach and Gleser (41) considered general distance measures. Rulon, et al (42) extensively reviewed the problem of inferring group membership from multivariate data. The concensus of these papers was that the centour score method should be used for the purpose of assigning individuals. The centour score is a function of the ratio of the distance between a point and the centroid of a group to the dispersion of the group. The computer programs developed by Cooley and Lohnes (6) for multiple discriminant analysis provide opportunity for classification by either of two methods: (1) assignment to the group with which it has the lowest $\chi^2$ score, or (2) assignment to the group with which it has the highest probability of membership. If the groups have equal dispersions and frequencies of membership, the two rules give the same assignment. The $\chi^2$ score method as shown by Cooley and Lohnes (6) is a function of the group means and dispersions. The second method of assignment employs Bayes' strategy and is a function of the group dispersions and sizes. Miller (10) presents an excellent review of both methods.

Miller (10), in discussing classification, suggests that a test statistic, $\overline{P}$, proposed by Brier and Allen (43) be used to test whether one set of discriminant functions performs better than another. The test is based on observed and predicted probabilities of group

membership. Sanders (44) has shown that the $\overline{P}$ score measures both the sharpness and validity of predicted probabilities and that the difference in $\overline{P}$ scores may be used in a statistical test for a significant improvement of one system over another. The test statistic, a function of the $\overline{P}$ scores, is distributed as t.

Summary

Based upon this review of the literature on multiple discriminant analysis, the data generated in this research project will be subjected to both two- and three-group analyses using the computer programs developed by Cooley and Lohnes. Wilks' $\Lambda$ will be used to test the significance of group separation. Bartlett's test statistic will be used to test the equality of group dispersions. Differences in Mahalanobis' $D^2$ will be used to test the significance of variables as they are added. Wallis' use of principal components analysis and varimax rotation in the selection of variables will be compared to Miller's stepwise selection of variables. The number of significant functions will be determined by Bartlett's test on root size.

Classification of the individuals will be used to analyze the groups and their membership. The centour score and Bayes' strategy will be used for assignment of observations to groups. The test of Brier and Allen will be used to check the validity of both the test for significance of variables and the test for the number of significant functions.

## Mathematical Models of Watershed Response

A large number of rainfall-runoff models have been developed, many in the last few years. Many of these models are concerned with

the distribution of rainfall excess in the form of a hydrograph of flow. However, this research is based on the volume of storm runoff, rainfall excess, therefore, some of the well known watershed models such as those developed by TVA, Crawford and Linsley, and parts of the Soil Conservation Service models are not discussed.

Many methods have been developed and proposed for calculating the volume of runoff or precipitation excess. The first chapter of reference (45) in the Literature Cited, a recent publication, is devoted to the calculation of rainfall excess. Since the chapter describes the state of the art so well, it has been abstracted in the following discussion of rainfall excess. Almost all of the methods that have been proposed can be classified as belonging to one or another of the following techniques:

The Bookkeeping Method or Threshold Concept

In agriculture, especially in irrigation, the bookkeeping method has been used for many years to estimate rainfall excess. In its simplest form it may be represented by a single reservoir with a given capacity. The water level in this reservoir declines continually between rainfall events due to evapotranspiration and this deficit must be replenished before runoff begins.

deZeeuw (46) made the storage capacity of the reservoir a function of the maximum deficit such that it could vary from year to year.

Makkink and van Heemst (47) divided the reservoir into three zones; i.e., the evaporation zone, the transition zone, and the ground water zone. They made the evaporation a function of the water present

in the evaporation zone.

Kohler (48) developed a model similar to that of Makkink and van Heemst in that it has two zones or reservoirs. The upper reservoir, which represents the upper layer of the soil profile, loses moisture at potential evapotranspiration rate. The lower level loses water only when the upper level is dry.

van Schilfgaarde (49) in studying drainage, used a model similar to both that of Makkink and van Heemst and that of Kohler. It is based on the assumption that the soil does not hold moisture above field capacity and that moisture moves downward only when the soil above it is at field capacity. Water for evapotranspiration is assumed to come from the surface layer. There is no restriction on the number or thickness of the layers as they develop.

The bookkeeping method as described in these examples is strongly related to physical soil properties and when one considers a drainage basin as a whole, there are a great variety of soil conditions which need to be considered independently. Kohler and Richards (50) developed a multicapacity accounting method which represents a basin by several separate reservoirs, each having its own maximum capacity. Within each reservoir, evaporation is set equal to potential evaporation until the soil moisture is depleted.

Graphical Correlation Methods

In these methods less attention is paid to the physical runoff process and more is paid to the factors that obviously influence the rainfall/rainfall excess relation. Graphical correlation techniques are used to quantify the factor.

Linsley, Kohler, and Paulhus (51) describe the methods used in graphical correlation using such factors as an antecedent precipitation index, storm duration, time of the year, and rainfall intensity.

Becker (52) used the coaxial method to develop a rainfall-runoff relation using an antecedent precipitation index, week of the year, ground water level, rainfall duration, and rainfall amount. In plotting the coaxial relations, he took into account theoretical considerations as to the direction and limits of the lines.

Infiltration Approaches

The principle of the infiltration approach is that runoff occurs when precipitation intensity exceeds the infiltration rate. Horton (53), promoter of the infiltration concept, derived his infiltration equation from plot studies. Estimating the initial infiltration rate, and compensating for periods within a storm when intensity is less than the infiltration rate are two difficulties arising in the use of the method.

Holtan (54) developed a more general equation and avoided the problem of an adequate and continuous supply of water by expressing infiltration capacity as a function of potential storage. Holtan's equation is identical to that of Horton under certain conditions.

In all infiltration approaches, the problem remains that of estimating initial conditions.

Functional Relations

The methods grouped under this title are based on the concept that a functional relation between rainfall and rainfall excess can be established. Such methods require estimates of both coefficients and

parameters in order to calculate the runoff. The coefficients are usually obtained by correlation studies. Kohler and Richards (50) established a functional relation between rainfall and rainfall excess, and used their multicapacity accounting model to estimate parameters of the function at the beginning of precipitation.

The USDA Soil Conservation Service (55) assumes a somewhat similar curvilinear relation exists between precipitation and precipitation excess, however, an initial abstraction is assumed to take place before runoff begins.

Infiltration Approach and Threshold Concept

This method tries to eliminate the problem of estimating the initial infiltration rate by correlating infiltration rates with an index parameter such as soil moisture and then using a bookkeeping technique to keep a running account of the index parameter. Kohler (56) solved the problem by relating the initial infiltration capacity to soil moisture deficit which he predicted using the multicapacity accounting technique. Holton (54) took into account the other problem associated with the infiltration approach, that of adjusting infiltration rates for storm periods in which intensity is less than infiltration rate.

Functional Relations and Threshold Concept

Functional relations by themselves have a somewhat similar problem to that of the infiltration approach, that of estimating values of parameters that are components of the system at the time just prior to the event. The bookkeeping technique is used to keep a running account of index parameters that can be correlated with variables

in the equation.

de Zeeuw (46) developed a functional relationship to distinguish between rainfall excess occurring as overland flow and rainfall excess occurring as ground water runoff. The volume of overland flow is subtracted from rainfall and the volume retained is used as input in the water balance model described earlier.

Wiser and van Schilfgaarde (57) used the Soil Conservation Service functional relationship in a somewaht similar manner to calculate the volume of runoff. Wiser was able to relate the curve number of the SCS model to the moisture content of the upper layer of the soil profile and could, therefore, use the SCS model in conjunction with their accounting model.

Summary

The watershed model to be selected for use must be capable of predicting storm runoff. In general, the bookkeeping techniques have not been satisfactory by themselves for this purpose. They have, however, been satisfactory for keeping a running account of parameters in various watershed models. Graphical correlation techniques are satisfactory for predicting a few events manually; however, in situations where a large number of events must be calculated, a computer oriented approach must be used. Both the infiltration and functional approaches are weak by themselves because of the necessity of estimating initial conditions.

On the basis of these statements and the material presented in this section of the report, it would be advisable to use either an infiltration or a functional relation combined with a bookkeeping

technique. Provision should also be made to provide for a rainfall threshold below which no runoff would be experienced.

## Generation of Rainfall Sequences

Sequences of rainfall have been used in studies of water resources systems for many years. The sequences can be divided into categories depending upon the time interval used. In the following survey of the literature distribution characteristics of annual, monthly, and daily rainfall are described. Because rainfall amounts for time intervals less than one day are not needed in this study, the characteristics of hourly data are not discussed. However, three good references on this subject are Pattison (58), Ramaseshan (59) and Grace and Eagleson (60). The report by Grace and Eagleson has an excellent review of the literature in this field and parts of the following are abstracted from it.

### Annual Rainfall

Many probability distributions have been fitted to annual rainfall amounts. Slade (61), using a logarithmic transformation of the normal distribution, was the first to fit a continuous distribution. Thom (62) and Merraim (63) drew a series of curves rather than fitting a specific distribution. Markovic (64) tried fitting five different distributions; a 2-parameter normal, a 2-parameter log-normal, a 3-parameter log-normal, a 2-parameter Gamma, and a 3-parameter Gamma. He found that the 2-parameter log-normal worked best in fitting his data. Results of these and other studies would indicate that the 2- or 3-parameter log-normal curve is best for annual data.

Meteorological observations are usually not independent of preceding conditions and their effects tend to carry over or persist into later times. The statistical significance of this tendency, known as "persistence," is evaluated by the serial correlation coefficient. Studies by Yule (65) in Britain, Kotz and Neumann (66) in Israel, Brittan, et al (67) in the mountainous area of western United States, Hoel (68) on the west coast of the United States, and Pattison (69), all would indicate that no significant persistence exists in annual rainfall data.

Monthly Rainfall

Many of the distributions which have been fitted to annual rainfall amounts have also been fitted to monthly rainfall data. Whitcomb (70) fitted a Gamma or Pearson Type III curve, whereas Stidd (71) and Beals (72) found that monthly as well as annual and daily precipitation, when raised to a fractional power were normally distributed. Other studies tend to substantiate these results and indicate that no one type of distribution has been found universally acceptable for monthly data.

Persistence in monthly as well as daily rainfall data was found to be significant by Besson (73,74) in the area near Paris. However, Beer, et al (75) could not find significant persistence in monthly rainfall in Britain. Namias (76) found that in the United States persistence was present in some but not all pairs of months.

Daily Rainfall

In the study of daily rainfall amounts, Das (77) and Kotz and Neumann (78) found the data to be adequately described by a Gamma

distribution; whereas, Brakensiek (79) was able to use a log-normal probability distribution. Beals (72) found that daily amounts raised to the one-fourth power were normally distributed.

Persistence in successive daily rainfall was found to be significant by Besson (73,74) in the vicinity of Paris, by Uttinger (80) in northern Italy, by Hannan (81) in Australia, and by Sellers (82) on the east coast of the United States. Williams (83), Longley (84), and Cooke (85) analyzed persistence on the basis of runs of dry or wet days rather than serial correlation and also found it to be significant. All of these studies would indicate that persistence in successive daily rainfall amounts should be present, however the distribution of daily amounts does not appear to be adequately defined by any one distribution function.

Models for Generating Daily Rainfall Sequences

Synthetic sequences of annual monthly and daily rainfall have generally been obtained by fitting the historic record to a probability distribution and then sampling this distribution with random numbers. If, however, persistence is found to exist in the historic record of a station being used as a model for development of a synthetic data generator, then the persistence must be included in the model by a serial correlation coefficient.

It was found in the previously discussed studies that annual and, in some cases, monthly precipitation amounts were uncorrelated with the previous month's amounts; i.e., little or no persistence, therefore generation of synthetic sequences is relatively easy. The generation of daily data is however not as simple because the tendency

for persistence appears to be quite strong.

In the following discussion, the generation of synthetic sequences has been broken into two components. The first is the occurrence or nonoccurrence of an event, and the second is the size of an event when it is generated.

Occurrence or Nonoccurrence of Daily Rainfall. Several different methods have been proposed for estimating the distribution of events. They have, however, been either a form of Markov process based on transition probabilities or a probability distribution fitted to the lengths of wet and dry periods.

Gabriel and Neumann (86) found that sequences of daily rainfall could be described by a first-order Markov process. The same system was found to be satisfactory at several locations by Caskey (87) and Weiss (88). Other authors, Newnham (89), Jorgensen (90), and Cooke (85) have not been as successful. Wiser (91) found that discrepancies occurred so consistently that he proposed a more general probability model of which the Markov chain model was a special case. Feyerherm and Bark (92) suggested the use of a second-order Markov chain model. They also suggested that the matrix of transition probabilities should include variation with time of the year. Green (93) used an exponential distribution for the lengths of wet and dry spells and sampled alternately from them. Grace and Eagleson (60) found that the distribution of lengths of wet and dry periods could be represented by a Weibull distribution.

Depths of Daily Rainfall. Most of the methods used to calculate daily rainfall amounts are based on regression type techniques with

stochastic components. This is because of the persistence element which has been found to be statistically significant in most studies. Both Enger (94) and Hammerle (95) used autoregressive techniques for predicting rainfall depths. Sellers (82) used the same technique but related rainfall only to the amount on the previous day. Pattison (58) tested two different models for predicting rainfall amounts. In the first model he selected rainfall amounts from probability distributions of observed data, taking into account transition probabilities in selecting the distribution. The second model was a linear regression model between consecutive hourly rainfall depths. Ramaseshan (59) working primarily with hourly data from annual maximum storms tested five different regression models with stochastic time intervals. Grace and Eagleson (60) used a linear regression equation in which storm depth was a function of storm duration. A probabilistic element was added to the intermediate and higher storm amounts.

Summary

Historic data available for the watershed selected for modeling will be analyzed to find the probability distribution which best fits the size distribution of daily (storm) events. The data will also be analyzed for persistence and correlation with other meteorological characteristics. The occurrence or nonoccurrence of an event will be calculated using either a Markov process with transition probabilities or by fitting the distribution of lengths of wet and dry periods to a probability distribution. The model selected for calculating the size of the events will probably be a regression type depending upon the persistence factor in the rainfall.

# CHAPTER III

# MULTIPLE DISCRIMINANT ANALYSIS

## Introduction

The concept of multiple discriminant analysis is perhaps best presented in an illustrative sense. Assume that a counselor at a college has a small group of students who would like to know what professional group their personality traits and characteristics would indicate they are most nearly like. If this counselor were familiar with multiple discriminant analysis, he would group previous graduates from the college into G different professions such as engineering, teaching, law, medicine, etc., selecting only those who were practicing in one of the fields and discarding the others. The counselor would then get the results of entrance examinations, aptitude tests, and any other information of a similar nature which was available on all students. Batteries of tests designed for counseling purposes may even be available.

The object of multiple discriminant analysis is to use these "test" scores to distinguish between the different professions. This is done by finding linear combinations of the P test scores which best separate the groups. These linear combinations, which are similar in appearance to multiple regression equations, would be used with the test scores from the group of interested students to calculate their discriminant scores. These discriminant scores would be compared with

medians of the G groups and then considering the group dispersions, the students could be told which of the G groups they are most nearly like.

In a more abstract sense, discriminant analysis is a procedure for looking at a number of groups from a direction that best separates the groups. The position of any element is defined by a series of linear functions, discriminant functions, all mutually orthogonal 1/. The maximum number of functions is the lesser of the two numbers G-1 and P.

A geometric interpretation of discriminant analysis may present the idea in a clearer manner. Consider a two-group two-test condition similar to the previous example where the groups are; A - engineering and B - teaching, and the tests are entrance examination results in X - mathematics and Y - English. The bivariate plot for the two groups, A and B, is shown in Fig. 1. The test scores for one individual, S, place him as shown in the figure. The two tests, X and Y, are correlated as shown by the major axes of the ellipses. Each ellipse, called a centile contour in the figure, represents the locus of points of equal density for a particular group. For example, the outer ellipse in each of the two groups might define the region within which 90 percent of the group lies.

The two points at which corresponding centours intersect define a straight line II. If a second straight line, I, is drawn perpendicular to line II, and if the points in the two-dimensional space are projected onto I, the overlap between the two groups will be smaller

---

1/ Orthogonal in the sense that the discriminant scores are uncorrelated

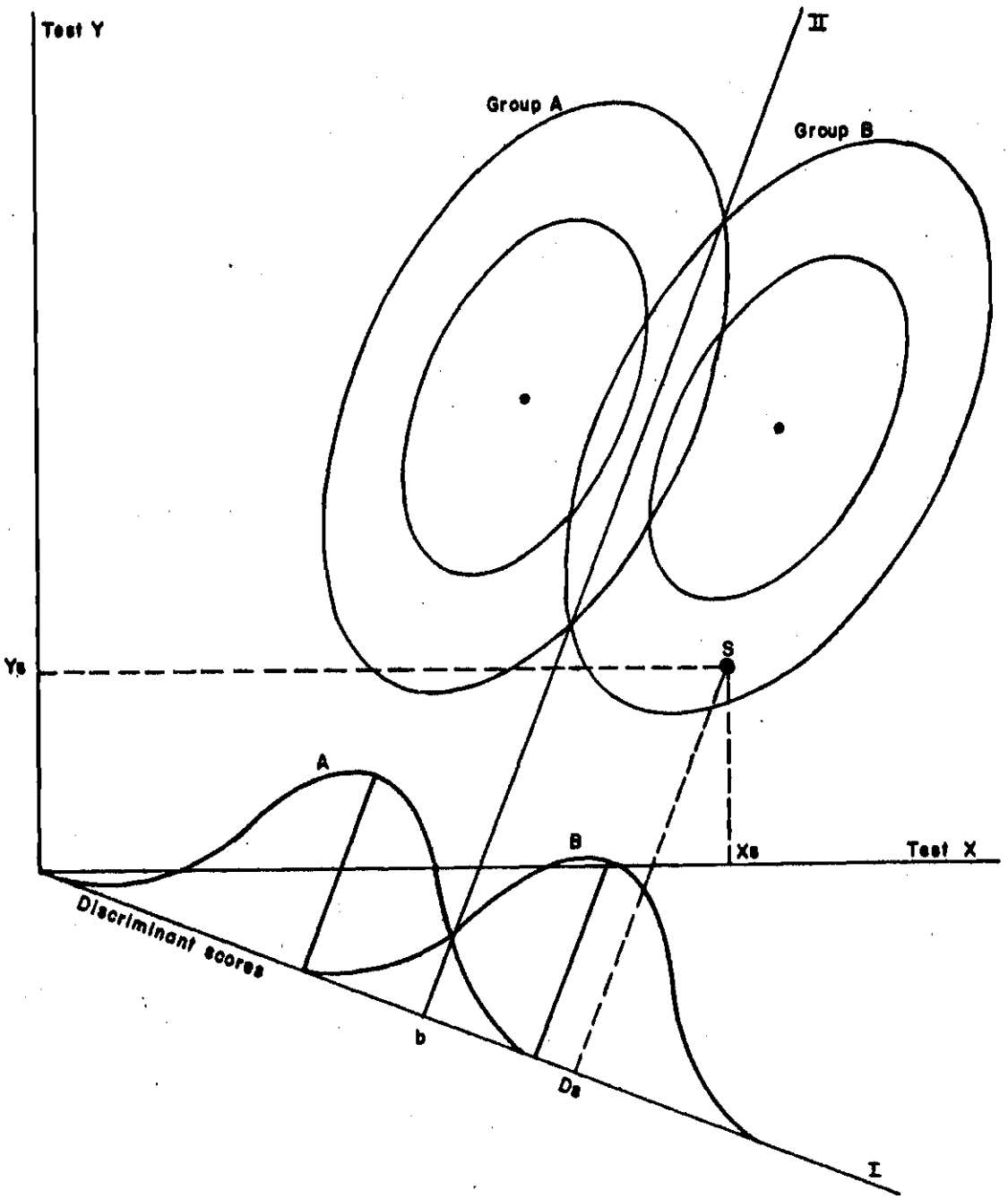



Figure 1. Geometric Interpretation of Discriminant Analysis

than for any other possible line. The discriminant function transforms the elements test scores to a single discriminant score which is its location along line I. For example, individual S's discriminant score is shown on line I as a function of test scores $X_s$ and $Y_s$. The point b divides the discriminant space into two regions indicating most probable membership in either group A or B.

## Mathematical Derivation

Given P test scores or predictors $X_p$ ($p = 1, \cdots, P$) representing measurements on G mutually exclusive and exhaustive groups with $n_g$ ($g = 1, \cdots, G$) being the number of elements or observations within each group, the $k^{th}$ individual element in the $g^{th}$ group on the $p^{th}$ predictor would be $X_{pgk}$. The total number of elements is

$$N = \sum_{g=1}^{G} n_g. \tag{1}$$

Consider a system of linear functions of the P predictors

$$Y_{jgk} = V_{j1} X_{1gk} + V_{j2} X_{2gk} + \cdots + V_{jp} X_{pgk} \tag{2}$$

$$(k = 1, \cdots, n_g)$$

$$(g = 1, \cdots, G)$$

$$(j = 1, \cdots, \min (G-1, P))$$

with the following characteristics:

The group mean of Y for the $j^{th}$ subsystem and the $g^{th}$ group is

$$\bar{Y}_{jg} = \frac{1}{n_g} \sum_{k=1}^{n_g} Y_{jgk}. \tag{3}$$

The sum of squares of Y between groups for the $j^{th}$ subsystem is

$$SSB(Y_j) = \sum_{g=1}^{G} n \, (\bar{Y}_{jg.} - \bar{Y}_{j..})^2. \tag{4}$$

The sum of squares of Y within groups for the $j^{th}$ subsystem is

$$SSW(Y_j) = \sum_{g=1}^{G} \sum_{k=1}^{n_g} (Y_{jgk} - \bar{Y}_{jg.})^2. \tag{5}$$

To maximize discrimination between the groups, the j sets of vector weights $V_{jp}$ (p=1, ····, P) are established so that the ratios $\lambda_j$ of the between-groups sum of squares to the within-group sum of squares are a maximum.

$$\lambda_j = \frac{SSB(Y_j)}{SSW(Y_j)} \tag{6}$$

The j values of $\lambda$ are indicative of a distance dimension of the subspace defined by the group means. Maximization is accomplished by expressing the sums of squares as quadratic forms in the predictor variate and then applying the ordinary techniques of differential calculus.

The matrix of sums of squares of Y between groups, B, and the sums of squares of Y within groups, W, used to calculate $\lambda$ are expressed in terms of the element X by using matrix notation as follows:

$$B = \begin{vmatrix} SSB(X_1) & SPB(X_2X_1) & \cdots & SPB(X_{p-1}X_1) & SPB(X_pX_1) \\ SPB(X_1X_2) & SSB(X_2) & \cdots & SPB(X_{p-1}X_2) & SPB(X_pX_2) \\ \vdots & \vdots & & \vdots & \vdots \\ SPB(X_1X_{p-1}) & SPB(X_2X_{p-1}) & \cdots & SSB(X_{p-1}) & SPB(X_pX_{p-1}) \\ SPB(X_1X_p) & SPB(X_2X_p) & \cdots & SPB(X_{p-1}X_p) & SSB(X_p) \end{vmatrix} \quad (7)$$

where B is the symmetrical between-group deviation sums of squares and products matrix.

$$W = \begin{vmatrix} SSW(X_1) & SPW(X_2X_1) & \cdots & SPW(X_{p-1}X_1) & SPW(X_pX_1) \\ SPW(X_1X_2) & SSW(X_2) & \cdots & SPW(X_{p-1}X_2) & SPW(X_pX_2) \\ \vdots & \vdots & & \vdots & \vdots \\ SPW(X_1X_{p-1}) & SPW(X_2X_{p-1}) & \cdots & SSW(X_{p-1}) & SPW(X_pX_{p-1}) \\ SPW(X_1X_p) & SPW(X_2X_p) & \cdots & SPW(X_{p-1}X_p) & SSW(X_p) \end{vmatrix} \quad (8)$$

where W is the symmetrical pooled within-group deviation sums of squares and products matrix. Elements of the matrices are defined as

$$SSB(X_p) = \sum_{g=1}^{G} n_g \, (\bar{X}_{pg.} - \bar{X}_{p..})^2 \quad (9)$$

$$(p = 1, \cdots, P)$$

where $SSB(X_p)$ is the sum of squared deviations between group means and

the grand mean for predictor $X_p$.

$$SPB(X_pX_q) = \sum_{g=1}^{G} n_g \, (\bar{X}_{pg.} - \bar{X}_{p..})(\bar{X}_{qg.} - \bar{X}_{q..}) \qquad (10)$$

$$(p, q = 1, \cdots, P; \; p \neq q)$$

where $SPB(X_pX_q)$ is the sum of products of deviations between group means and the grand mean for predictors $X_p$ and $X_q$ $(p, q = 1, \cdots, P;$ where $p \neq q)$.

$$SSW(X_p) = \sum_{g=1}^{G} \sum_{k=1}^{n_g} (X_{pgk} - \bar{X}_{pg.})^2 \qquad (11)$$

$$(p = 1, \cdots, P)$$

where $SSW(X_p)$ is the pooled within-group sum of squared deviations about the group means for predictor $X_p$.

$$SPW(X_pX_q) = \sum_{g=1}^{G} \sum_{k=1}^{n_g} (X_{pgk} - \bar{X}_{pg.})(X_{qgk} - \bar{X}_{qg.}) \qquad (12)$$

$$(p, q = 1, \cdots, P; \; p \neq q)$$

where $SPW(X_pX_q)$ is the pooled within-group sum of products of deviations about the group means for predictors $X_p$ and $X_q$, $(p, q = 1, \cdots, P;$ where $p \neq q)$.

The $j^{th}$ column vector of predictor weights is defined as

$$V_j = \begin{vmatrix} V_{j1} \\ V_{j2} \\ \cdot \\ \cdot \\ \cdot \\ V_{jp-1} \\ V_{jp} \end{vmatrix} \tag{13}$$

Using matrix notation and algebra, Bryan (8) shows that the two sums of squares of Y as defined by equations 4 and 5 may be written respectively as

$$V'BV \text{ and } V'WV \tag{14}$$

where $V'$ is the transpose of $V$. The general form of Equation 6, the ratio of the between to the within sums of squares, is therefore

$$\lambda = \frac{V'BV}{V'WV} \tag{15}$$

Bryan (8) also shows that by setting the partial derivatives of $\lambda$ with respect to $V_1, V_2, \cdots, V_n$ equal to zero, the following matrix equation is obtained:

$$[V'WV]\ BV - [V'BV]\ WV = 0. \tag{16}$$

By dividing through by $V'WV$ and collecting terms

$$[B - \lambda W] V = 0 \tag{17}$$

or

$$[W^{-1}B - \lambda I] V = 0 \tag{18}$$

where $W^{-1}$ is the inverse of W, and I is the identity or unit matrix. The $W^{-1}B$ matrix is nonsymmetric.

The individual ratios $\lambda_j$ are roots of the determinantal equation

$$[W^{-1}B - \lambda I] = 0 \tag{19}$$

and the vectors $V_j$ are solution of equation 18 with $\lambda$ equal to $\lambda_j$, i.e.,

$$[W^{-1}B - \lambda_j I] V_j = 0 \tag{20}$$

$$(j = 1, \cdots, \min (G-1, P)).$$

The vectors $V_j$ from Equation 20 are the weights which define the discriminant functions given by

$$Y_j = V_{j1}X_1 + V_{j2}X_2 + \cdots + V_{jp}X_p \tag{21}$$

$$(j = 1, \cdots, \min (G-1, P))$$

for any set of observations $X_p$ $(p = 1, \cdots, P)$. Equation 21 is the same as Equation 2 except that it applies to a specific element in a specific group rather than the general application given by Equation 2.

## Multivariate Statistical Tests

### Hotelling's $T^2$

Hotelling's $T^2$ is used in discriminant analysis to infer whether or not a group of predictors contributes a significant amount of information to discriminate between two groups. It tests in particular the null hypothesis that there is no significant difference between the mean vectors $\bar{X}_{pe\cdot}$ and $\bar{X}_{pf\cdot}$ for the two groups e and f considering the dispersions within the groups. The test statistic, $T^2$, is

$$T^2 = \left| \frac{\bar{n}_e + n_f - \bar{2}}{\frac{1}{n_e} + \frac{1}{n_f}} \right| d' W_{ef}^{-1} d \tag{22}$$

where $n_e$ and $n_f$ are the number of observations in groups e and f, respectively, $W_{ef}^{-1}$ is the inverse of the pooled within-group deviation sums of squares and products matrix for groups e and f only, and d and d' are the vector and its transpose of the difference in means between groups e and f for the p predictors.

$$d = \begin{bmatrix} (\bar{X}_{1e\cdot} - \bar{X}_{1f\cdot}) \\ (\bar{X}_{2e\cdot} - \bar{X}_{2f\cdot}) \\ \vdots \\ (\bar{X}_{(p-1)e\cdot} - \bar{X}_{(p-1)f\cdot}) \\ (\bar{X}_{pe\cdot} - X_{pf\cdot}) \end{bmatrix} \tag{23}$$

The null hypothesis for which $T^2$ is the test statistic is that

the expected value of vector d is zero. The test is set up as

$$\frac{n_e + n_f - P - 1}{(n_e + n_f - 2)\ P} T^2 \sim F(P,\ n_e + n_f - P - 1). \tag{24}$$

Wilks' Λ

Wilks' Λ is the multi-group extension of Hotelling's $T^2$ and is used to test the overall discriminating power of a group of predictors. As with the $T^2$ statistic, the within-group dispersions are considered. The test stistic, Λ, is

$$\Lambda = \left|\frac{W}{T}\right| \tag{25}$$

where W is the pooled within-group deviation sums of squares and products matrix and T is the total sample deviation sums of squares and products matrix equal to the sum of the W and B matrices

$$T = W + B \tag{26}$$

where B is the between-groups deviation sums of squares and products matrix.

Rao (3) showed that Λ could be tested by the F distribution following a transformation. The following quotation from Cooley and Lohnes (6) outlines the procedure:

Let $s = \sqrt{(p^2q^2 - 4)/(p^2 + q^2 - 5)}$, $q = g - 1$

$m = n - (p + q + 1)/2$, $n = N - 1$

$\lambda = -(pq - 2)/4$,

$r = pq/2$,

and $y = \Lambda^{1/s}$;

then

$$(4.1) \qquad F_{ms + 2\lambda}^{2r} = \left|\frac{\bar{I} - \bar{y}}{\bar{y}}\right| \left|\frac{\overline{ms + 2\lambda}}{\overline{2r}}\right|$$

For the one-variate case, this F transformation of $\Lambda$ is the algebraic equivalent of the familiar univariate F test. This is because $T = A + W$ where $A$ is the usual among groups sums of squares. The F test can be written

$$F = \left|\frac{\bar{A}}{\bar{W}}\right| \left|\frac{\bar{N} - \bar{g}}{\bar{g} - \bar{1}}\right|$$

which becomes (recalling $A = T - W$)

$$F = \left|\frac{\bar{I} - (\bar{W}/\bar{T})}{W/T}\right| \left|\frac{\bar{N} - \bar{g}}{\bar{g} - \bar{1}}\right|$$

or

$$F = \left|\frac{\bar{I} - \bar{\Lambda}}{\Lambda}\right| \left|\frac{\bar{N} - \bar{g}}{\bar{g} - \bar{1}}\right|$$

Equation 4.1 reduces to this last equation for the univariate case $p = 1$.

In the above quotation p and g have the same meaning as has been used previously, i.e., p is the number of predictors or variables, and g is the number of groups. The ratio, $\lambda$, in the quotation is not the same as $\lambda$ in the previous discussion. The A matrix is equivalent to the B matrix described previously.

The null hypothesis, $H_2$, for which $\Lambda$ is the test statistic is

$\mu_1 = \mu_2 = \cdots = \mu_g$ where the $\mu_s$ are population centroids for the g groups. The feasibility of a test of $H_2$ is based on the assumption that population dispersions, variance covariance matrices, $\Delta$'s, are equal. Thus a second hypothesis $H_1 : \Delta_1 = \Delta_2 = \cdots = \Delta_g$, that the dispersion matrices are equal and from common dispersions, is postulated.

Testing the Equality of Dispersion Matrices. Cooley and Lohnes (6) present the following test, attributed to Bartlett (35) and Box (36), for testing the null hypothesis $H_1$:

Box defines the criterion M:

$$M = n \log_e |D| - \sum_g (N_g \log_e |D_g|)$$

where $D = 1/n\ W$, $D_g = (1/n_g)\ W_g$ and $n = N - g$. Required parameters are:

$$A_1 = \left[ \sum_g \frac{1}{n_g} - \frac{1}{n} \right] \frac{2p^2 + 3p - 1}{6(g - 1)(p + 1)}$$

$$A_2 = \left[ \sum_g \frac{1}{n_g^2} - \frac{1}{n^2} \right] \frac{(p - 1)(p + 2)}{6(g - 1)}$$

If $A_2 - A_1^2$ is positive, then

$$f_1 = .5(g - 1)p(p + 1), \quad f_2 = (f_1 + 2)/(A_2 - A_1^2),$$

$$b = f_1/(1 - A_1 - f_1/f_2),$$

$$F \begin{matrix} f_1 \\ f_2 \end{matrix} = M/b$$

If $A_2 - A_1^2$ is negative, use the following:

$$f_1 = .5(g - 1)p(p + 1), \quad f_2 = (f_1 + 2)/(A_1^2 - A_2),$$

$$b = f_2/(1 - A_1 + 2/f_2),$$

$$F \begin{matrix} f_1 \\ f_2 \end{matrix} = f_2 \, M/f_1 \, (b - M).$$

The tests of $H_2$ and $H_1$ are quite involved, therefore quite often the equality of dispersions is assumed and not tested. In many cases, this assumption may be justified as Cooley and Lohnes (6), pp. 61 and 66, point out that $H_2$ is rather insensitive to moderate departure from homogeneity of dispersions.

Mahalanobis' $D^2$

Mahalanobis' $D^2$ is a measure of the "distance" between two groups and is directly proportional to Hotelling's $T^2$. Rao (3) extended it to situations of more than two groups. For P predictors, it is defined as

$$D_P^2 = (n - 1 - GP) \text{ trace } W^{-1} B \tag{27}$$

where n, G, and P are respectively, the sample size, the number of groups, and the number of predictors. $W^{-1}$ and B have their usual meaning. The trace of a matrix is the sum of the diagonal elements of

the matrix. According to Rao (54), for a large number of observations $D_P^2$ is estimated as

$$D_P^2 \sim \chi^2 \ (P(G - 1)). \tag{28}$$

The number of statistically significant predictors used in the discrimination can also be tested using the $D_P$ statistic. The difference in $D^2$ with the addition of Q predictors over and above the original P is, for large n, approximately

$$(D_{P+Q}^2 - D_P^2) \sim \chi^2 \ (Q(G - 1)) \tag{29}$$

where

$$D_{P+Q}^2 = [n - 1 - G(P + Q)] \text{ trace } W_{P+Q}^{-1} B_{P+Q}$$

with $W_{P+Q}^{-1}$ and $B_{P+Q}$ the same as W and B but including all P+Q variables or predictors.

Use of $D^2$ in the Stepwise Selection of Predictors. Miller (10) using the distribution represented by Equation 29 developed a stepwise approach to the selection of predictors. It is analogous to the stepwise selection of variables in multiple regression analysis. He uses as the criterion, $D^2$, and the test for selection of the $S^{th}$ selected predictor as

$$(D_S^2 - D_{S-1}^2) > (G - 1)\chi^2_{\alpha^*/(P-S+1)} \qquad (30)$$

$$(S = 1, \cdots, r).$$

where $\alpha^*$ is the probability that one or more of these predictors are judged significant when, in fact, none of the P predictors is significant. He shows that division of $\alpha^*$ by $(P - S + 1)$ is a valid adjustment to compensate for the fact that as predictors are selected, the opportunity of selection by chance increases.

The first variable is selected by calculating the trace $W^{-1} B$ for each of the predictors and then selecting the predictor X(1) which has the largest trace, i.e.

$$\text{trace } W^{-1} B \ (X^{(1)}) \geq \text{trace } W^{-1} B \ (X_p) \qquad (31)$$

$$p = (1, \cdots, P).$$

It is tested using Equation 30 which becomes

$$(D_1^2 - D_0^2) > (G - 1)\chi^2_{(\alpha^*/P)} \qquad (32)$$

where $D_0^2 = 0$. If Equation 32 is satisfied, the trace $W^{-1} B \ (X^{(1)}X_p)$ $(p = 1, \cdots, P; X_p \neq X^{(1)})$ is evaluated for each of the remaining $P - 1$ predictors and again the largest combination selected is tested using Equation 30. The procedure is continued until R predictors have been selected such that the $R + 1$ selection does not satisfy Equation 30.

Testing the Significance of Roots of $W^{-1} B$

Miller (10) presents two methods attributed to Bartlett and Rao

(3) of judging the statistical significance of roots, $\lambda_j$, of the $W^{-1}$ B matrix.

$$[n - 1 - 1/2\ (P + G)]\ \ln\ (1 + \lambda_j) \sim \chi^2\ (P + G - 2j) \qquad (33)$$

The alternate method which is less refined but equivalent for large n is

$$(n - 1 - G)\ \lambda_j \sim \chi^2\ (P + G - 2j). \qquad (34)$$

In both equations n is the sample size and j has a limit of G - 1 or P whichever is smaller. Equations 33 and 34 when used to test the significance of the roots of the $W^{-1}$ B matrix for the R selected predictors show a significant root if

$$[n - 1 - 1/2\ (R + G)]\ \ln\ (1 + \lambda_j) > \chi^2_{(\alpha^*/t-j+1)}\ (R + G - 2j) \qquad (35)$$

or, using the alternate method, if

$$(n - 1 - G)\ \lambda_j > \chi^2_{(\alpha^*/t-j+1)}\ (R + G - 2j) \qquad (36)$$

where t is G - 1 or R whichever is smaller and j is the number of the root selected and has a limit of t. The significance level of $\chi^2$, $\alpha^*$, is adjusted for probability of selection in a manner similar to the $\chi^2$ adjustment in the test for significant predictors. See Miller (10).

## Classification

### Introduction

Discriminant analysis which leads to the development of the discriminant function is primarily concerned with testing null hypotheses and studying group differences in terms of group mean vectors, group dispersions, adjusted group means, or configuration of group centroids in the discriminant space. The logical next step is to examine an element to see which group it is most like. This is the problem to which classification is most closely associated. It has been used primarily in the fields of taxonomy, career guidance, and the selection and assignment of manpower. It does, however, also give insight into the relationship between an individual element and the groups.

### The Mathematics of Classification

The centour score has been proposed (41,6) as the best method for assessing the degree to which an element resembles each of several groups in terms of a set of predictor variables. In the two-variate case, the centour or centile contour is an ellipse in two dimensional space, $X_1$, $X_2$, within which a certain percent of the elements are expected to lie (see Fig. 2). The ellipse is defined by its $\chi^2$ value.

$$\chi^2(2) = x_i' D^{-1} x_i \qquad (37)$$

where $\chi^2$ is distributed with 2 degrees of freedom, $D^{-1}$ is the inverse of the dispersion matrix and $x_i$ and $x_i'$ are the vector, and its transpose, of deviations of a point i to the centroid of the group (see Fig. 2).

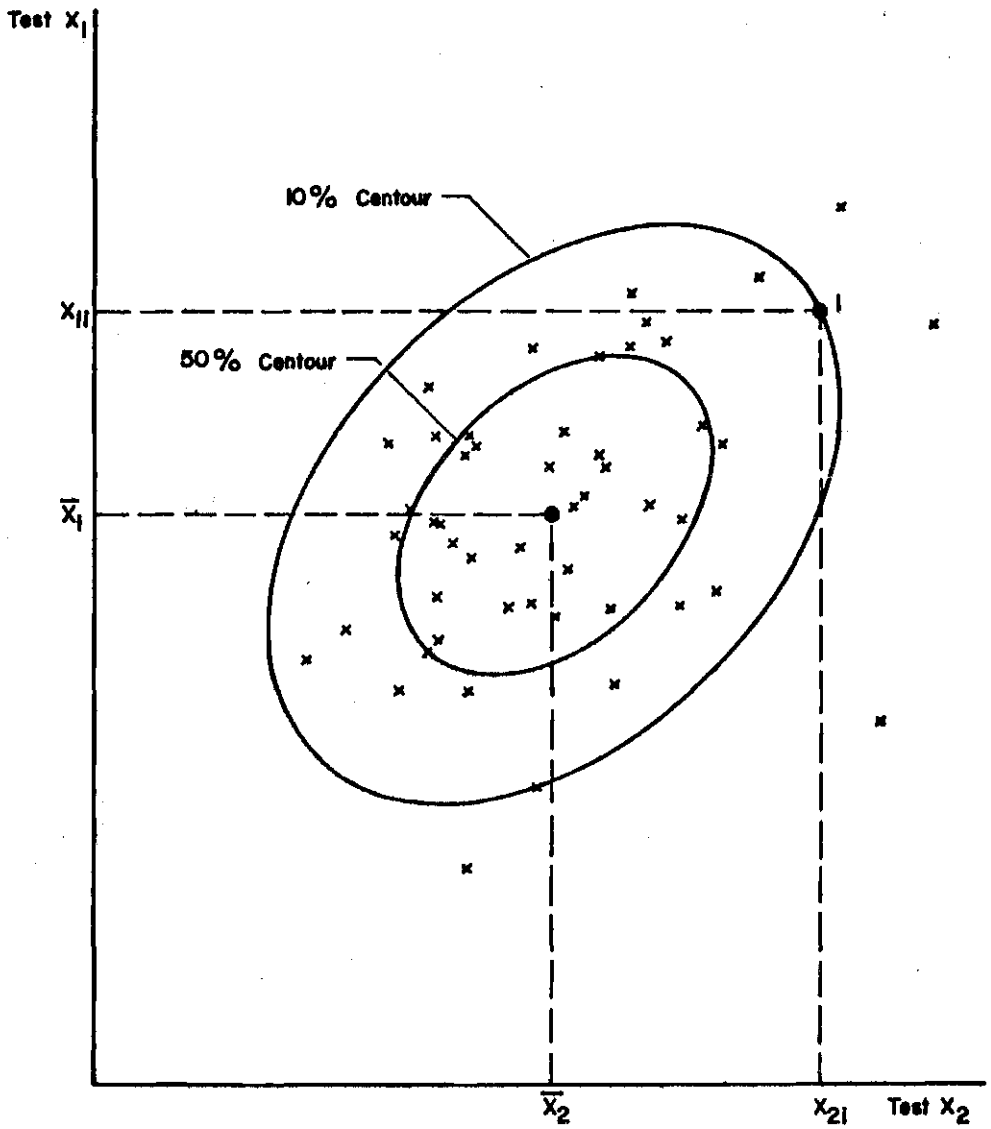



Figure 2. Centours of a Group of Fifty Observations on Each of Two Tests - $X_1$ and $X_2$

$$X_i = \begin{vmatrix} \bar{X}_{1i} - \bar{X}_1 \\ X_{2i} - \bar{X}_2 \end{vmatrix} \tag{38}$$

where $X_{1i}$ and $X_{2i}$ are values on each of the 2 variables for a point i. Suppose for example Equation 37 were to give a $\chi^2$ value of 4.61 for point i. Entering the $\chi^2$ table with 2 degrees of freedom, we find that the probability of lying further from the centroid than point i is 0.10. This 10 percent centour and the 50 percent centour, i.e., $\chi^2 = 1.39$, are shown on Fig. 2.

In the P variate case, the centours become hyperellipsoids with $X_i$ defined as

$$X_i = \begin{vmatrix} \bar{X}_{1i} - \bar{X}_1 \\ X_{2i} - \bar{X}_2 \\ \vdots \\ X_{pi} - \bar{X}_p \end{vmatrix} \tag{39}$$

The problem of considering more than one group can best be seen by again considering the bivariate case with the two groups shown in Fig. 1. If point S represents an element from one or the other of the two groups, its $\chi^2$ values could be calculated with respect to each group. It is quite obvious from the location of S that it would lie on a mugh higher centour (lower $\chi^2$ score) of group B than of group A and would therefore be more likely a member of group B than group A. Thus the centour score can be used to assign an element to a group.

Using the centour or $\chi^2$ score as the only criterion for assignment, a decision rule can be postulated as follows:

$$\text{Rule I} \qquad R_g \cap \chi_g^2 \leq \chi_k^2 \qquad (40)$$

$$(g, k = 1, \cdots, G;\ g \neq k).$$

The rule states: The region of the test space for group g ($R_g$) is defined as ($\cap$) the space for which the group, g, $\chi^2$ is smaller than any other group, k, $\chi^2$. If the group dispersion matrices are equal and if the number of observations in each group are equal, this decision rule will result in a minimum number of misclassifications.

If, in addition to the centour score, size of the group and group dispersion are to be included in the criteria for assignment, then probability of group membership can be computed by using Bayes' theorem, Cooley and Lohnes (6),

$$P_{ig}\ (H_g|X_i) = \frac{\dfrac{P_g}{|D_g|^{1/2}}\, e^{-\frac{\chi_g^2}{2}}}{\sum_k \dfrac{P_k}{|D_k|^{1/2}}\, e^{-\frac{\chi_k^2}{2}}} \qquad (41)$$

$$(k = 1, \cdots, g, \cdots, G)$$

$$(i = 1, \cdots N)$$

Using the probability of assigning element i to group g as calculated from Equation 41, an alternate decision rule based on

probabilities can be postulated

$$\text{Rule II} \qquad R_j \cap P_{ig} \geq P_{ik} \qquad (42)$$

$$(g, k = 1, \cdots, G;\ g \neq k).$$

Using this rule, an element is assigned to the group for which its probability of membership is highest.

Discriminant Space

In the discussion so far, the location and assignment of an element has been with reference to what might be called the test space, i.e., based on the element's test scores. It is possible, however, to reduce the demensionality of the problem by working in a reduced space, i.e., discriminant space. In many cases, this can be a saving in time and cost and in some instances, it may be necessary in order to get the problem to a workable size. The centroid, dispersions, and element scores in reduced space are calculated from the discriminant functions. The matrix C of the centroids of the G groups in reduced, discriminant space is as follows:

$$C_{(T,G)} = V'_{(T,R)} \cdot M_{(R,G)} \qquad (43)$$

where T is the number of statistically significant roots, R is the number of predictors selected, and G is the number of groups. V is the matrix of T discriminant functions made up of R predictors and M is the matrix of G group means on each of the R predictors in the test space.

The matrix $DD_g$ of the group dispersions in reduced, discriminant

space is:

$$DD_{g(T,T)} = V'_{(T,R)} \cdot D_{g(R,R)} \cdot V_{(R,T)} \qquad (44)$$

where T, R, G, and V are as defined above, and $D_g$ is the R by R dispersion matrix for group g in the test space.

The discriminant scores $S_g$ for the elements in group g in reduced discriminant space is

$$S_{g(n_g,T)} = X_{g(n_g,R)} V_{(R,T)} \qquad (45)$$

where T, R, and V are as defined above, $n_g$ is the number of observations in group g, and $X_g$ is the matrix of $n_g$ element test scores on the R predictors in group g.

Using Equations 43, 44, and 45, the data may be reduced from the test space to the discriminant space at a savings in dimensionality of the problem. Cooley and Lohnes (6) make the following statement with reference to the use of data in the reduced space rather than the test space:

> An individual who lies in Region $R_j$ in the test space will also lie in $R_j$ in the discriminant space, if group dispersion matrices are equal. Therefore, under these conditions, the parameters estimated in equation 7.2 may describe the discriminant space rather than the test space, thus saving an enormous amount of computing time when the number of discriminant functions is substantially smaller than the number of tests, and the number of individuals to be classified is large (as is generally the case).
>
> Experience seems to indicate that moderate departure from homogeneity of dispersion does not produce differences between test space and discriminant space results.

In the above statement his region $R_j$ is the same as $R_g$ in Equations 40 and 42 and his Equation 7.2 is the same as Equation 41.

Test on the Precision and Validity of a Probability Prediction System

Miller (10) presents work by Sanders (44) and Brier and Allen (43) in which a test statistic $\overline{P}$ is proposed for measuring the precision and validity of a probability prediction system and can be used to compare two systems. The $\overline{P}$ score which is a function of the probabilities of group membership of an element under two different prediction systems, is defined as

$$\overline{P} = \frac{1}{M} \sum_{m=1}^{M} \sum_{g=1}^{G} (\tilde{P}_{gm} - O_{gm})^2 \qquad (46)$$

where M is the total number of elements, G is the number of groups, $\tilde{P}_{gm}$ is the calculated probability of the $m^{th}$ element being in the $g^{th}$ group, and $O_{gm}$ takes on a value of unity or zero depending upon the true group membership. The score has a maximum of two and a minimum of zero, where the lower values represent more desirable probabilities.

Following is a quotation from Miller describing the use of the $\overline{P}$ score in comparing two prediction schemes.

> although the sampling distribution of $\overline{P}$ has not been investigated, it is possible to perform a t test on the difference between $\overline{P}$ scores to determine if the probabilities of one probability prediction system is significantly superior to another. That is, for two probability prediction systems, say A and B,

$$\frac{(\bar{P}_A - \bar{P}_B)\cdot\sqrt{M(M-1)}}{\sqrt{\sum_{m=1}^{M}(P_{Am} - P_{Bm})^2 - \frac{\left[\sum_{m=1}^{M}(P_{Am} - P_{Bm})\right]^2}{M}}} \sim t\,(M-1) \qquad (57)$$

under the null hypothesis that the difference between the true means is zero and assuming the sample mean differences are normally distributed. The terms $P_{Am}$ and $P_{Bm}$ are the computed values for a single observation m in (56) of the probability prediction system A and B, respectively, and their means over all M observations are denoted as $\bar{P}_A$ and $\bar{P}_B$.

In the above quotation, the referenced equation 56 is the same as Equation 46 in this Chapter.

## Computer Programs for Multiple Discriminant Analysis

Computer programs for carrying out the analyses and tests described in this Chapter have been developed by several people. The programs used in the analyses and tests reported in the remainder of the dissertation were those developed by Cooley and Lohnes and described in excellent detail in their book (6). The programs were used essentially as they were presented with modifications required to fit them on an IBM 1130 with one disc. In addition to these changes, the program for calculating the discriminant vectors and $\Lambda$ was changed to include the calculation of Mahalanobis' $D^2$. The programs are so nearly like those of Cooley and Lohnes that they are not further described nor presented in this dissertation.

## Significance Level of Hypotheses Testing

The level of significance used is assessing the hypotheses presented in this chapter and throughout the remainder of the dissertation

was 0.05. All tests of significance were based on a Type I error, i.e. a low level of probability that a hypothesis H is rejected which should have been accepted. At no point in the dissertation is the significance level of a Type II error evaluated. A Type II error is that made in accepting a hypothesis which should have been rejected.

# CHAPTER IV

## THE WATERSHED AND ITS MODEL

Watershed D at the Blacklands Experimental Watershed near Riesel, Texas (see Fig. 3) was selected for modeling in this study because it had the following characteristics:

(1) It is located in a subhumid to semi-arid part of the country.

(2) It is large enough to have mixed land usage.

(3) It has a long enough period of record to emphasize each of the major land uses, and has satellite subwatersheds from which the hydrologic characteristics of the different land uses can be determined.

(4) It has a period of record long enough to cover the normal range in climatic conditions.

### The Watershed

The watershed, 1,110 acres in size, is located in the Blacklands of Central Texas in the Brazos River basin. The soils, montmorilonitic in nature, are deep, fine-textured, slowly-permeable residuals of marl and are subject to extensive and deep shrinkage cracks caused by drying. Land use varies from year to year but consists primarily of cultivated crops, sowed crops, permanent grass, roads, and farmsteads. Table 1 shows the land use from 1937 through 1966. Land use for the period from 1943 through 1948, the war and post war years, is not know. All land uses were grouped into five categories based on hydrologic

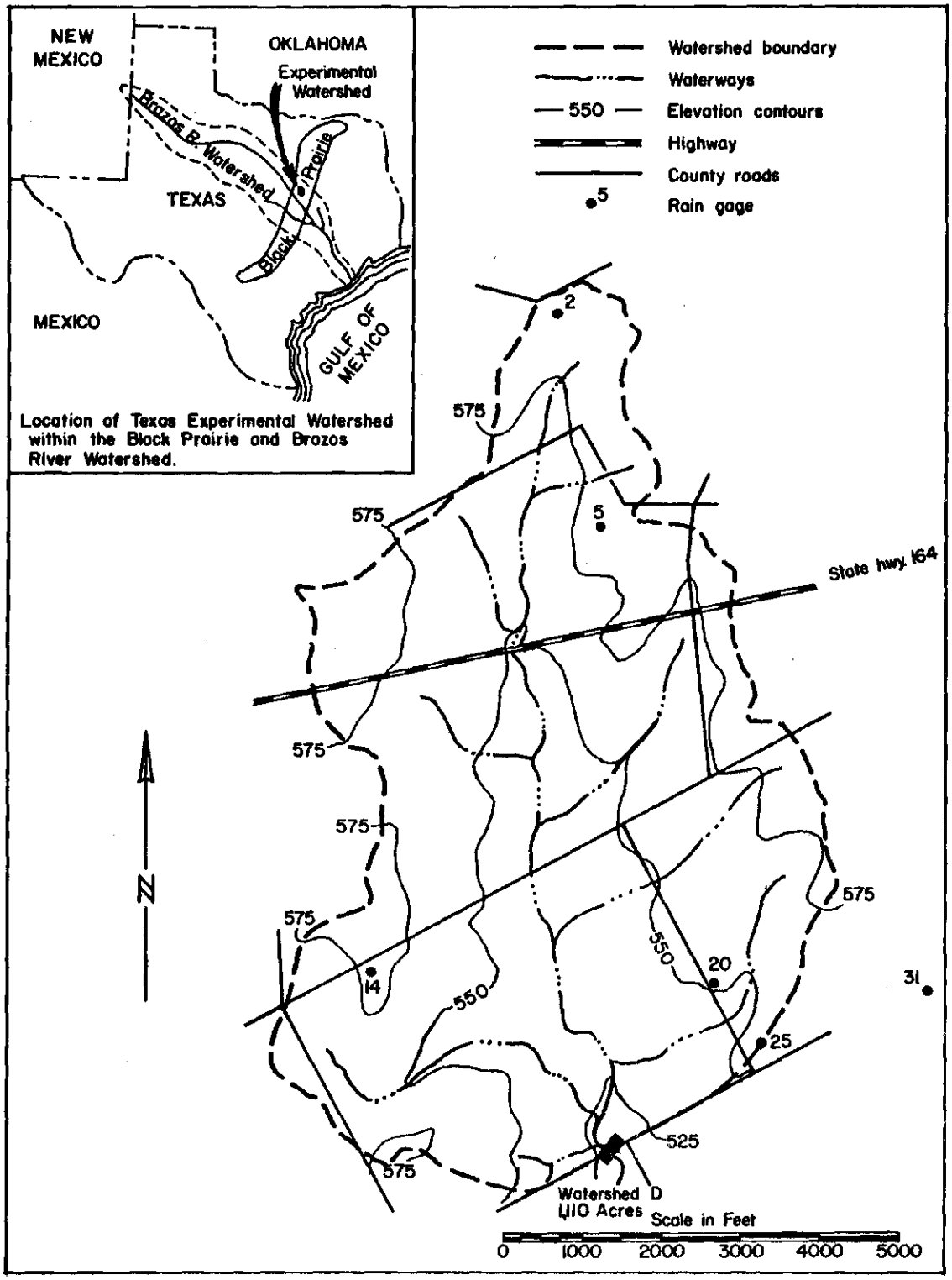



Figure 3. Blacklands Experimental Watershed, Riesel, Texas

similarity. Roads and farmsteads were classified as cultivated-no crops.

Table 1. Land Use for Watershed D, Riesel, Texas

| Year(s) | Land use expressed as a fraction of the area | | | | |
|---|---|---|---|---|---|
| | Bermuda Pasture | Native Grass Meadow | Cultivated Row Crops | Cultivated Oats | Cultivated No Crops |
| 1937 | .110 | .090 | .720 | .013 | .067 |
| 1938 | .111 | .142 | .648 | .012 | .087 |
| 1939 | .110 | .112 | .637 | .027 | .114 |
| 1940 | .108 | .096 | .658 | .019 | .119 |
| 1941 | .110 | .090 | .520 | .030 | .250 |
| 1942 | .140 | .090 | .600 | .020 | .130 |
| 1949-1957 | .250 | .000 | .639 | .037 | .074 |
| 1958 | .225 | .239 | .401 | .090 | .045 |
| 1959 | .207 | .272 | .351 | .122 | .048 |
| 1960 | .206 | .454 | .220 | .051 | .069 |
| 1961 | .205 | .485 | .212 | .026 | .072 |
| 1962 | .218 | .464 | .161 | .088 | .069 |
| 1963 | .448 | .308 | .136 | .056 | .052 |
| 1964 | .522 | .188 | .174 | .077 | .039 |
| 1965 | .470 | .162 | .190 | .140 | .038 |
| 1966 | .688 | .077 | .148 | .060 | .027 |

Watershed runoff is ephemeral, occurring only in direct response to rainfall. The watershed lies in an area where the volume of surface runoff per unit area does not generally change with size of the watershed. Average annual rainfall is about 33 inches with over one-third occurring in April, May, and June. It comes primarily in thunderstorms of high intensity and short duration. A more detailed description of the physiography, geology, soils, climate and agricultural practices is given in Hydrologic Bulletin No. 5 (96).

Data Available

A continuous record of streamflow from the watershed is available from November 12, 1937 to the present time. The station was

closed, however, for a period beginning during World War II from July 1, 1943 to March 1, 1949. The station is a current-meter station with an artificial low-water control. The stage is recorded on a 6-hour FW-1 chart. Rainfall for the watershed was estimated from the rain gages shown on Fig. 3. These weighing, recording gages were in operation at different times. The Thiessen weights for the rain gages are shown along with their effective dates in Table 2.

Table 2. Thiessen Weights of Rain Gages in the Watershed

| Date | Rain Gage Number and Thiessen Weight | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2 | 5 | 14 | 20 | 25 | 26A | 31 |
| Nov. 12, 1937 to Dec. 31, 1938 | | 35.21 | 51.44 | | | | 13.35 |
| Jan. 1, 1939 to Dec. 31, 1941 | 26.03 | | 15.06 | | | | 58.91 |
| Jan. 1, 1942 to March 31, 1942 | | 35.21 | 51.44 | | | | 13.35 |
| April 1, 1942 to July 1, 1943 | | 34.33 | 41.22 | | 24.45 | | |
| March 1, 1949 to Aug. 30, 1957 | | 35.72 | 35.89 | 28.39 | | | |
| Sept. 1, 1957 to Present Time | | 35.02 | 35.44 | 28.70 | | 0.84 | |

Other climatic data collected at the station include relative humidity, wind movement, temperature, and pan evaporation. Four different types of evaporation pan were used during the period of record at the station. Table 3 shows the effective dates of these pans.

Table 3. Pan Evaporation Record at Riesel, Texas

| Effective Date | Type of Evaporation Pan | | | |
|---|---|---|---|---|
| | Young's Screen | Colorado | BPI | USWB |
| Oct. 1938 | | X | | |
| Nov. 1938 | | | | X |
| Dec. 1938-Jan. 1939 | | X | | |
| Feb. 1939-Sept. 13, 1950 | | | X | |
| Sept. 14, 1950-Sept. 28, 1950 | | X | | |
| Sept. 29, 1950-June 1954 | | | X | |
| July 1954-May 11, 1960 | X | | | |
| May 12, 1960-May 24, 1960 | | X | | |
| May 25, 1960-Present | X | | | |

## The Watershed Model

On the basis of the literature survey it was concluded that the watershed model should be either an infiltration or a functional relation combined with a bookkeeping technique with provision for a rainfall threshold below which no runoff would be experienced. It was found that a model had already been developed for the area by M. A. Hartman (97,98). His model would be classified as a functional relationship combined with a threshold concept. It was therefore selected to reproduce the hydrologic characteristics of the watershed. Runoff as defined by the model is a function of storm rainfall, pan evaporation, soil moisture, and land use. For a given land use, the runoff is assumed to have a hyperbolic functional relationship with rainfall.

$$\frac{P}{P-Q} = a + bP \tag{47}$$

i.e., the ratio of rainfall to water retained by the soil is a linear

function of rainfall.

The functional form of the relationship provides for an initial abstraction of rainfall, $P_1$, before runoff begins. This initial abstraction is highly correlated with the antecedent soil moisture, and a linear function was used to express the relationship.

The slope factor b in Equation 47 was also correlated with the antecedent soil moisture by a set of linear equations. The parameter, a, is calculated from Equation 47 noting that at the instant runoff begins, Q is zero and P is $P_1$. The equation, therefore, reduces to

$$1 = a + b\ P_1 \tag{48}$$

The index of soil moisture used in the relationship with $P_1$ and b has been defined as the amount of soil moisture in excess of 18 percent in the top three feet of the soil profile. It was found that a continuous record of the soil moisture index could be obtained by a bookkeeping technique with additions to the soil moisture being the storm water retention, P-Q, and depletions being the estimated evapotranspiration. Evapotranspiration losses are approximated by the soil moisture depletion rate, k in the equation

$$SM_t = SM_o\ k^t \tag{49}$$

where $SM_o$ and $SM_t$ are the soil moisture values at times t days apart. The depletion rate, k, was found to be a linear function of the initial soil moisture, $SM_o$, and the average daily pan evaporation for the time

interval, t.

The rainfall-runoff relation as a function of antecedent soil moisture is shown for native grass meadow in Fig. 4.

Equations for calculating b, $P_1$, and k for each of the five land use conditions in Table 1 were developed from data on watersheds three acres in size located in the same general area as watershed D. The multiple correlation coefficients of the equations for calculating these relations were statistically significant at the one percent level. The equations for calculating the depletion rate, k; the initial abstraction, $P_1$; and the parameter, b, are presented in Tables 4, 5, and 6, respectively.

The following description shows how this model, developed by Hartman, was fitted to watershed D.

## Fitting the Model to the Watershed (Calculating the Probabilistic Element)

In setting up the model to generate data for discriminant analysis, it was desired that it be representative of many typical models used in project design. But at the same time, it should incorporate all the variation of a true watershed so that the significance of changes in land use parameters can be examined in comparison to the unexplained variance.

The unexplained variance of the model is the difference between the observed and predicted events and is made up of two elements which cannot be separated. Part of the unexplained variance is caused by errors in the watershed data. However, most of the unexplained

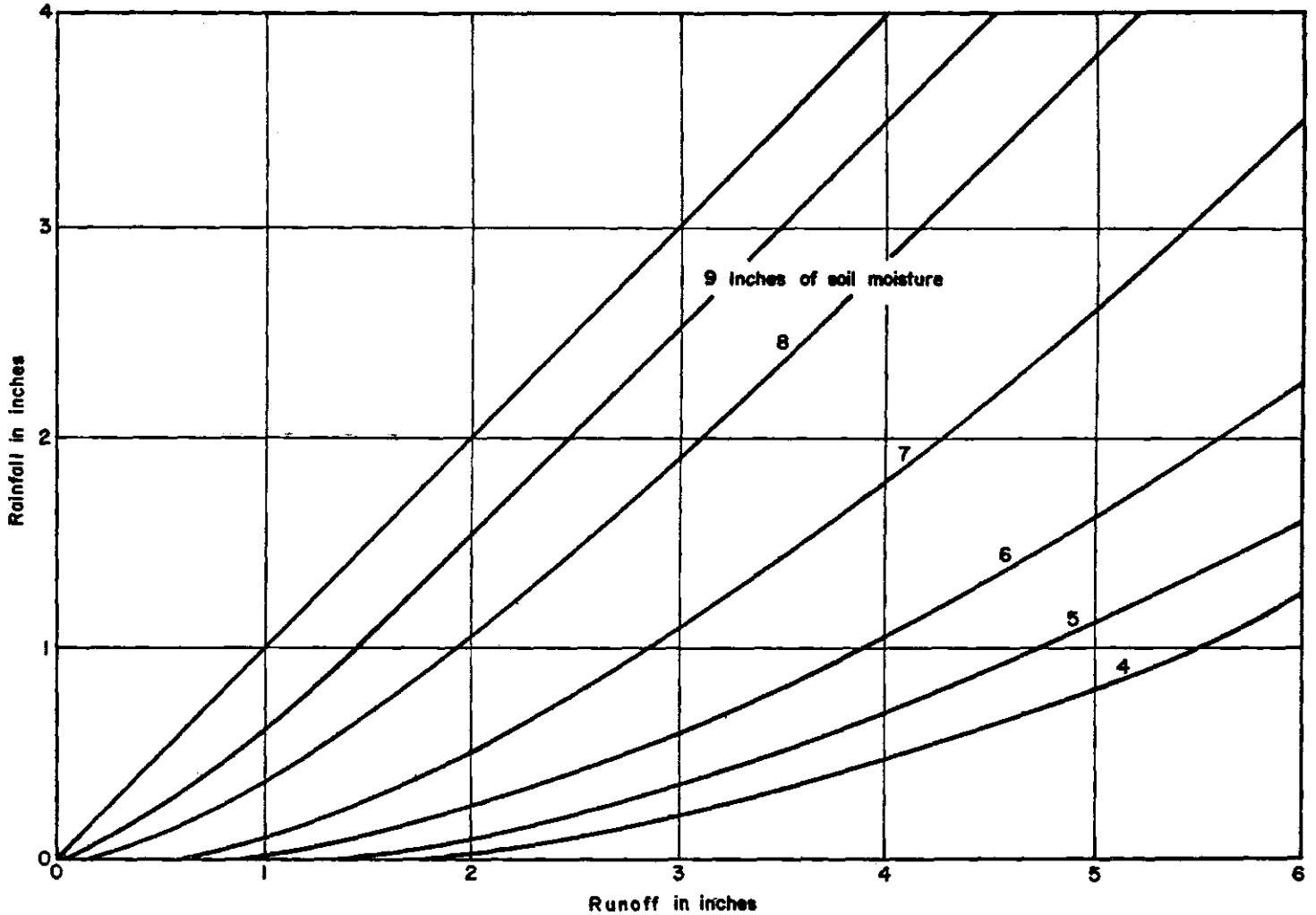



Figure 4. Rainfall-Runoff Relation as a Function of Antecedent Soil Moisture for Native Grass Meadow at Riesel Texas

Table 4. Equations for Calculating the Depletion Constant

$$K_{SM} = 0.982 + 0.005\ SM_o - 0.289\ PE$$
$$K_{WM} = 0.992 + 0.002\ SM_o - 0.145\ PE$$
$$K_{SP} = 0.958 + 0.003\ SM_o - 0.012\ PE$$
$$K_{WP} = 0.990 + 0.001\ SM_o - 0.033\ PE$$
$$K_{SO} = 0.967 + 0.001\ SM_o - 0.009\ PE$$
$$K_{WO} = 0.988 + 0.003\ SM_o - 0.167\ PE$$
$$K_{SNC} = 0.956 + 0.005\ SM_o - 0.029\ PE$$
$$K_{WNC} = 0.986 + 0.001\ SM_o - 0.019\ PE$$
$$K_{SRC} = 0.930 + 0.005\ SM_o - 0.060\ PE$$

$K_{SM}$ = native grass meadow depletion rate for March-October
$K_{WM}$ = native grass meadow depletion rate for October-March
$K_{SP}$ = Bermuda pasture depletion rate for March 1-October 15
$K_{WP}$ = Bermuda pasture depletion rate for October 15-March 1
$K_{WO}$ = cultivated-oats depletion rate for December-February
$K_{SO}$ = cultivated-oats depletion rate for March 1-May 15
$K_{SNC}$ = cultivated-no crop depletion rate for March 1-October 15
$K_{WNC}$ = cultivated-no crop depletion rate for October 15-March 1
$K_{SRC}$ = cultivated-row crop depletion rate for May 15-October 15
$SM_o$ = soil moisture in excess of 18 percent in top three feet of soil at end of last day's observation (inches)
PE = average daily pan evaporation for the period (inches-Young screen pan)

During period when no crop is growing, the no-crop rate is used.

Table 5. Equations for Calculating the Initial Abstraction

$$P_{im} = 3.37 - 0.41\ SM_o$$
$$P_{ip} = 1.39 - 0.16\ SM_o$$
$$P_{ic} = 1.98 - 0.22\ SM_o$$
$$P_{io} = 3.06 - 0.37\ SM_o$$
$$P_{in} = 1.98 - 0.22\ SM_o$$

$P_{im}$ = initial abstraction for native grass meadow
$P_{ip}$ = initial abstraction for Bermuda pasture
$P_{ic}$ = initial abstraction for cultivated-row crops
$P_{io}$ = initial abstraction for cultivated-oats
$P_{in}$ = initial abstraction for cultivated-no crops
$SM_o$ = antecedent soil moisture in excess of 18 percent which is assumed to be the inches of moisture in the top three feet of soil.

Table 6. Values of Coefficients $a_1$ and $b_1$ Used to Calculate the Parameter b

Basic Equation: $b = \frac{1}{b_1 + a_1 SM_o}$

| Land Use | SM Range | $a_1$ | $b_1$ |
|---|---|---|---|
| Bermuda pasture | <1.88 | 1.48 | 17.44 |
| " " | 1.88-5.55 | 3.21 | 20.69 |
| " " | >5.55 | 1.25 | 9.82 |
| Native grass meadow | <4.94 | 1.48 | 17.44 |
| " " " | 4.94-7.98 | 2.85 | 24.21 |
| " " " | >7.98 | .90 | 8.65 |
| Cultivated-row crops | <3.79 | 1.48 | 17.44 |
| " " " | 3.79-5.92 | 4.66 | 29.50 |
| " " " | >5.92 | .54 | 5.12 |
| Cultivated-oats | <1.55 | 1.48 | 17.44 |
| " " | >1.55 | 2.24 | 18.62 |
| Cultivated-no crop | <3.79 | 1.48 | 17.44 |
| " " " | 3.79-5.66 | 4.66 | 29.50 |
| " " " | >5.66 | .80 | 7.65 |

$SM_o$ is the antecedent soil moisture, in excess of 18 percent, which is assumed to be the inches of moisture in the top three feet of soil

variance is probably due to an incomplete description of the rainfall-runoff phenomena.

In order to develop a model such as the one just described, it must be made up of two components, (1) a deterministic element representative of a typical watershed model, and (2) a probabilistic element equal to the unexplained variance of the system. The unexplained variance of the system is determined by processing the observed rainfall and pan evaporation through the watershed model and comparing the runoff prediced with the observed.

Input Data. Storm rainfall for use in the model was defined as daily rainfall. Consecutive storms, i.e., two or more consecutive days of rainfall, were separated by at least 8 hours. If the storms were separated by less than 8 hours, they were combined and placed on the first day. In the period of record, very few storms were combined. In the area in which watershed D lies, most rainfall comes as thunderstorms in the late afternoon and lasts only a few hours.

The runoff from these storms lasted in most cases only a few hours and at most three or four days. When rainfall events were close enough that runoff from the first storm had not stopped, the runoff events were separated by using typical recession curves.

Pan evaporation as used in the model is that of the Young's screen pan. Therefore, conversion factors were needed to adjust evaporation from the other three pans listed in Table 3 to that of the Young's screen pan. The energy balance for each of the four pans is different because the pans are of different sizes and not exposed to wind action in the same way. Since evaporation is a function of both

the water temperature and wind action in addition to other factors, the four pans will have different rates of evaporation.

Ratios published in U.S. Geological Survey Professional Paper 269 (99) were used to convert the Weather Bureau and Colorado pans to that of a Young's screen pan. The ratios for converting the BPI pan data to a Young's screen pan were obtained from data published in Texas Agricultural Experiment Station Bulletin 787 (100). Pan evaporation data for both the Young's screen pan and the BPI pan were published for the period 1943 through 1953 for Buchanan dam in Texas. Ratios of the monthly data from both pans were averaged by months and the monthly ratios used to adjust the BPI records. The ratios for all three pans are presented in Table 7.

Table 7. Factors for Converting USWB, Colorado, and BPI Evaporation Pan Data to Young's Screen Pan

| Month | USWB | Colorado | BPI |
|---|---|---|---|
| January | .94 | .97 | .88 |
| February | .64 | 1.01 | 1.00 |
| March | .80 | .97 | 1.04 |
| April | .79 | .91 | 1.07 |
| May | .66 | .86 | 1.09 |
| June | .66 | .83 | 1.10 |
| July | .68 | .85 | 1.09 |
| August | .67 | .84 | 1.05 |
| September | .72 | .90 | 1.04 |
| October | .75 | .93 | 1.03 |
| November | .97 | 1.11 | .92 |
| December | 1.04 | 1.27 | .86 |

Evaporation data for the missing period in 1937 and 1938 were taken from records at Temple, Texas. A 13-year period, 1940-1953,

showed that the average evaporation from the BPI pan at Riesel was 1.33 times that at Temple. This ratio was used to adjust the daily records at Temple to those at Riesel. The adjusted values were then corrected to that of the Young's screen pan for use in the model.

Processing Data through the Model. The rainfall, evaporation, and land use for the period of record were processed through the watershed model. Soil moisture for use in the model was calculated from rainfall, evaporation and observed runoff. The difference between the calculated and observed runoff values is equivalent to the probabilistic element described earlier. Fig. 5 is a log-log plot of the calculated runoff, $q_c$, vs. the observed runoff, $q_o$.

The disparity between the observed and predicted values of Fig. 5 would lead one to question the adequacy of the model. However, if the data were replotted on arithmetic rather than log-log paper, the disparity would not appear as great. In the description of the Blacklands area it was mentioned that the soils were montmorilonitic, slowly permeable and subject to the development of deep cracks. Thus the soils can vary from almost completely impermeable when wet to almost completely open when dry. Another factor to consider is the fact that a 2-inch rain on a recently plowed field could easily be absorbed producing no runoff; whereas, the same field prior to plowing might produce 1 inch of runoff from the same rainfall. The watershed used in this investigation is small enough, 1,110 acres, that runoff from recently plowed fields could produce a significant change in runoff from the area. Yet the watershed is large enough that it would be nearly an impossible task to keep track of the cultural practices on

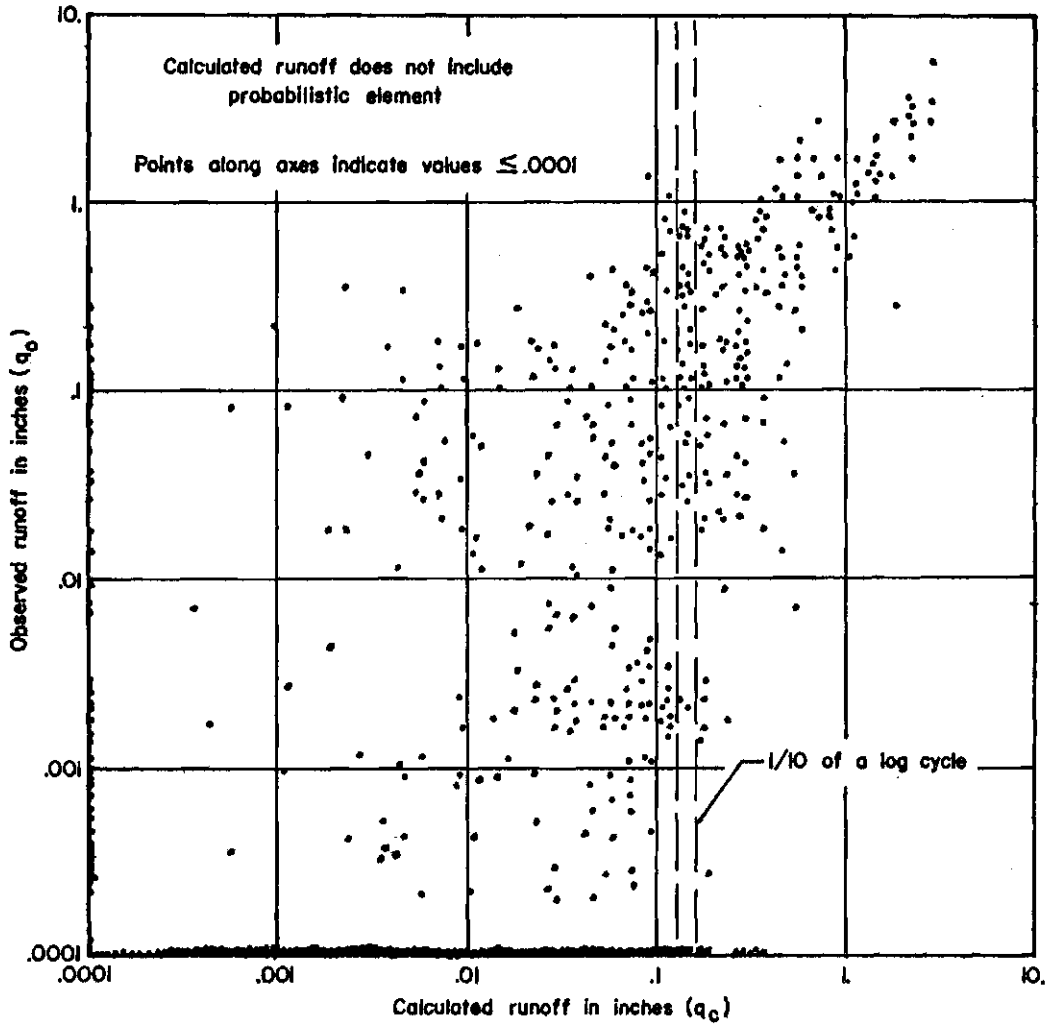



Figure 5. Calculated Runoff vs. Observed Runoff

each farm. Combining the effects of watershed size with the soil characteristics would thus help explain some of the disparity between the observed and calculated runoff.

Fig. 5 also shows that the runoff model predicts many more runoff events of 0.10 inch or less than were actually observed, and that the deviation between the observed and calculated data is not uniform with respect to the calculated value. The statistical properties of these deviations were used to develop the probabilistic element. They are described below.

The large number of events on Fig. 5 with no observed flow were eliminated by stratifying the calculated values into 10 uniform intervals per log cycle (see Fig. 5) and then linearly relating the percent of zero valued points in each stratum to the mean value of the stratum. Fig. 6 is a plot of this data. The two linear equations are:

$$P_z = 57.1 - 10.7 \log q_c \qquad (50)$$

$$q_c < 0.0195 \text{ inches}$$

and

$$P_z = -30.0 - 61.6 \log q_c \qquad (51)$$

$$q_c > 0.0195 \text{ inches}$$

Distribution of the remaining events in the plot on Fig. 5 represents variance in the model system which must be explained by the probabilistic element.

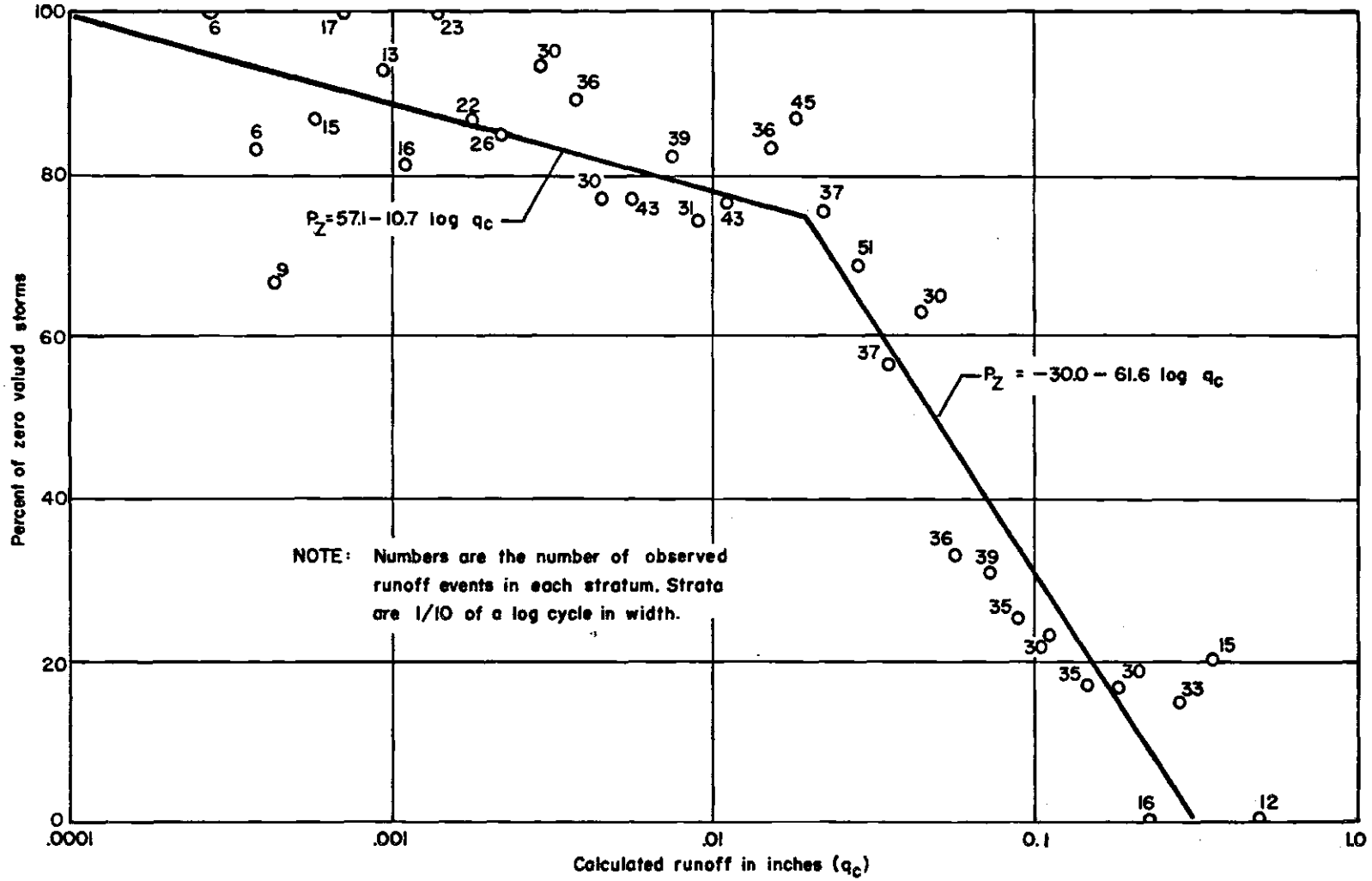



Figure 6. Percent of Zero Valued Points Per Stratum as a Function of the Calculated Runoff

The equation used in adding the probabilistic component is given by

$$q_d = q_c + \left| \overline{t_i} \; (s_\varepsilon) + \overline{\overline{\varepsilon}} \right| \tag{52}$$

in which $q_d$ and $q_c$ are the distributed and calculated values of runoff, respectively, $t_i$ is a standardized, random, normal, and independently distributed variate, $s_\varepsilon$ is the standard deviation of the residuals, and $\overline{\varepsilon}$ is the mean of the residuals.

A superficial study of Fig. 5 shows that a simple scheme such as represented by Equation 52 will not be satisfactory because the distribution of points about the line of equal values is not uniform.

Logarithms of observed runoff events in each stratum below 0.1 inch were found to be uniformly distributed over a very broad range. The distribution within each stratum is shown diagrammatically in Fig. 7. The lower limit of the uniform part of the distribution was 0.0001 inch for all strata. The upper limit of the uniform part of the distribution shown as $P_{ue}$ in Fig. 7 is a function of the calculated runoff. The upper limit of the distribution, shown in Fig. 7 as $P_{oe}$, is also a function of the calculated runoff. The equations representing these two limits were calculated from the distribution of points in Fig. 5 and are:

$$\log P_{ue} = 0.44 + 0.56 \log q_c \tag{53}$$

$$\log P_{oe} = 0.60 + 0.40 \log q_c \tag{54}$$

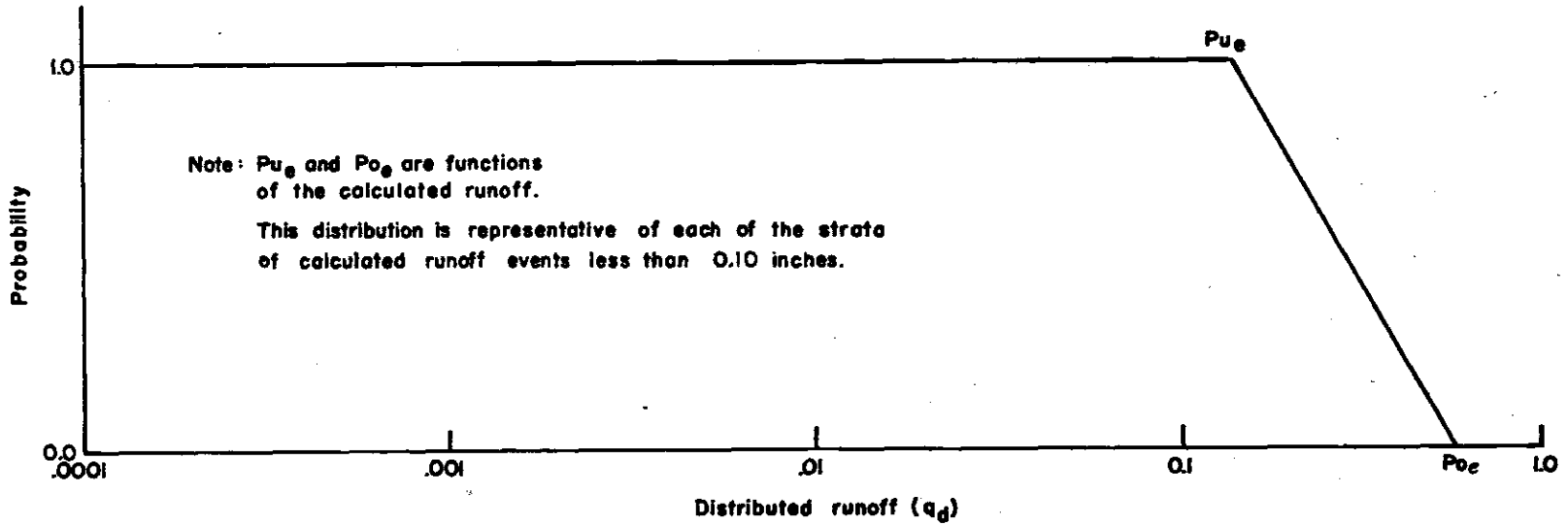



Figure 7. Probability Density Function of Observed Runoff Events within One Stratum

Since the observed events were nearly uniformly distributed for calculated events less than 0.1 inch, the distributed runoff becomes entirely probabilistic. The rejection technique, Grace and Eagleson (60), pp. A-25, 26, was used to sample the density function of Fig. 7. to obtain the distributed runoff from calculated runoff events less than 0.1 inch. Tests on the adequacy of the technique are presented further in the report.

The distribution of observed events within each stratum of calculated runoff greater than 0.1 inch as shown in Fig. 5 was investigated. The investigation showed that neither a normal nor a log-normal distribution was satisfactory. However, a skewed log-normal distribution adequately represented the wide variations in distribution form exhibited by the data in Fig. 5. Equation 52 was changed by noting that within any one stratum, $\overline{q}_o$, the mean observed runoff, is equal to the sum of $q_c$ and $\overline{\varepsilon}$. The logarithmic form of the equation is

$$\log q_d = \log \overline{q}_o + t_{i*}\, s_o \tag{55}$$

in which $\overline{q}_o$ is the observed mean at $q_c$, $s_o$ is the standard deviation of the logarithms of the observed runoff events at $q_c$, and $t_{i*}$ (see Reference 101) is a skewed, random, normal, and independently distributed variate, i.e.,

$$t_{i*} = t_i + 0.16\, g_o\, (t_i^2 - 1) \tag{56}$$

in which $g_o$ is the skew coefficient of the logarithms of the observed

runoff events at $q_c$. But in looking at each of the strata in Fig. 5 starting with the first one above 0.10 inch and going toward the right (higher calculated runoff values), it can be seen that the variance within each stratum gradually gets smaller. It is not possible to see what happens to the mean and skew coefficient by looking at Fig. 5, therefore the mean, $\overline{q}_o$, the standard deviation, $s_o$, and the skew coefficient, $g_o$, of the logarithms of the events in each of the strata were plotted vs. the logarithm of calculated runoff in Figs. 8, 9, and 10, respectively. This data shows that all three parameters are functions of the calculated runoff, $q_c$. Thus linear functions were fitted to the data. The equations fitted by least squares are shown on the figures. Two line segments were used to define the mean, $\overline{q}_o$, and three were used for the standard deviation, $s_o$. This is because data from strata below 0.1 inch showed that a broken line was more satisfactory than a single line. The least squares lines were fitted to the data by weighting the points by the number of observations they represent. Thus very little weight was given to the points with few observations. This approach is especially justified in this case because all three parameters are for practical purposes meaningless when based on as few as five or less observations.

The equations for calculating the log-mean observed event are:

$$\log \overline{q}_o = .130 + 1.810 \log q_c \qquad (57)$$

$$-1.0 \leq \log q_c \leq -.42$$

$$\log \overline{q}_c = -.035 + 1.416 \log q_c. \qquad (58)$$

$$-.42 \leq \log q_c$$

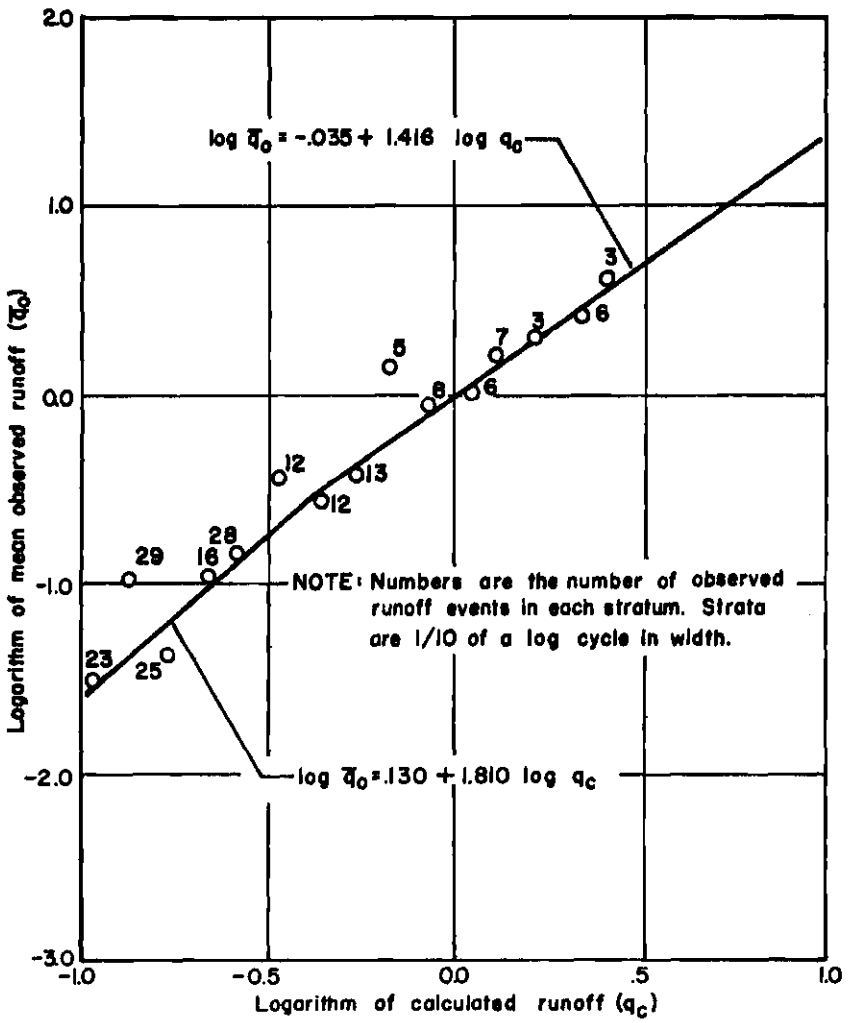



Figure 8. Mean Observed Runoff in each Stratum of Calculated Runoff

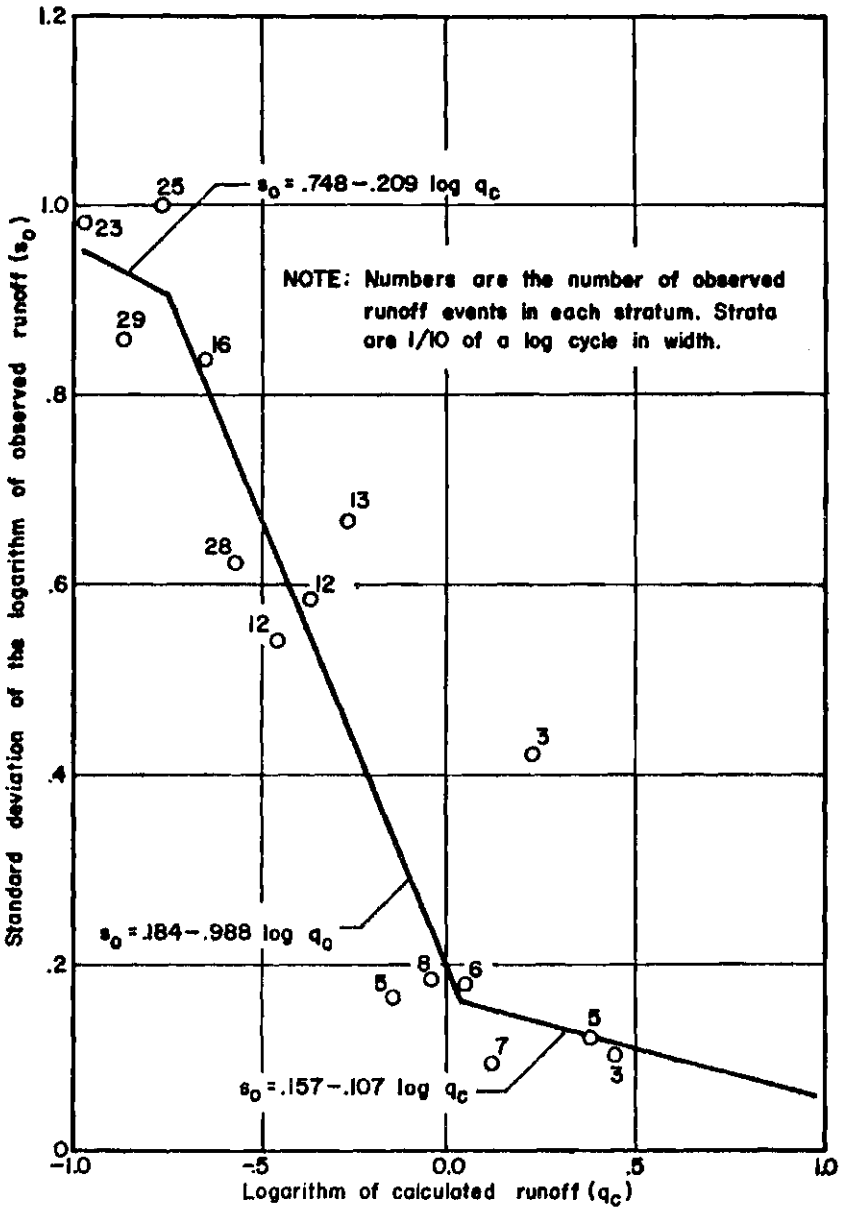



Figure 9. Standard Deviation of Observed Events in Each Stratum of Calculated Runoff

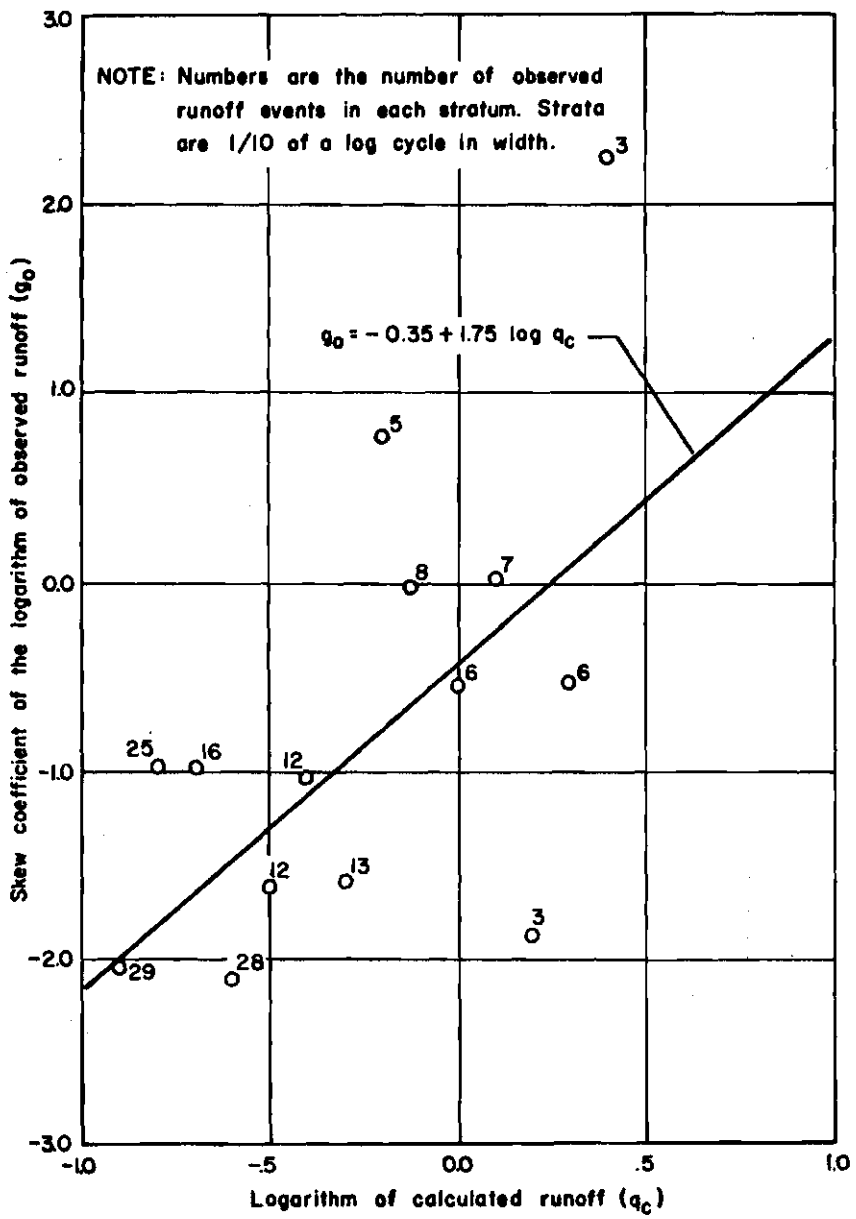



Figure 10. Skew Coefficient of Observed Events in each Stratum of Calculated Runoff

The equations for calculating the standard deviations of the logarithms of the events are:

$$s_o = .748 - .209 \log q_c \qquad (59)$$

$$-1.0 \leq \log q_c \leq -0.725$$

$$s_o = .184 - .988 \log q_c \qquad (60)$$

$$-0.725 \leq \log q_c \leq 0.030$$

and

$$s_o = .157 - .107 \log q_c. \qquad (61)$$

$$0.030 \leq \log q_c$$

The equation for calculating the skew of the logarithms of the events is

$$g_o = -0.35 + 1.75 \log q_c. \qquad (62)$$

$$-1.00 \leq \log q_c$$

Using these equations, the mean, $\overline{q}_o$, the standard deviation, $s_o$, and the skew coefficient, $g_o$, can be calculated from an event, $q_c$, for use in Equation 55. For example, suppose the watershed model were to give a calculated runoff event of 1.0 inch; Fig. 8 shows that the mean of a distributed or observed event for a calculated event of this size would be about 0.92 inch, Fig. 9 shows that the standard deviation would be about 0.184 inch, and Fig. 10 shows that the skew coefficient would be about -0.35.

Assuming further that the random normal deviate, $t_i$, for this event were +0.5, then substitution of $g_o$ and $t_i$ into Equation 56 would make $t_i*$ equal to 0.54. Using this value of $t_i*$ and the calculated values of $\overline{q}_o$ and $s_o$, Equation 55 gives the size of the runoff event including the probabilistic element as 1.08 inch.

In the previous discussion, logarithms of observed runoff in each stratum of calculated runoff rather than the arithmetic values were used for two reasons:

(1) Logarithms of the events were more nearly normally distributed than were arithmetic values.

(2) By working with logarithms, negative values of runoff would not develop when adding the probabilistic element.

The equations for incorporating the probabilistic element into the watershed model were added to the computer program previously described. They were tested by using the observed land use, and processing the rainfall and evaporation through the model. Two tests were made using different starting points in the random number generator. (See the Appendix, Chapter A-1, for a discussion of the random number generator.) The Kolmogorov-Smirnov two-sample test (see Appendix, Chapter A-2, for a discussion of this test) was used to evaluate the adequacy of the scheme.

Results of applying the test to the two synthetic runs are shown in Table 8. The maximum differences for the two runs were 4.2 and 3.8. The critical values of $D_n$ at the 5 percent level based on an observed record of 452 events and synthetic records of 391 and 441 were 9.4 and 9.1, respectively. These results indicate that the null hypothesis

cannot be rejected and that the two synthetic runs are from the same population as was the observed record.

A plot of the calculated runoff, $q_c$, vs. the distributed runoff, $q_d$, for one of the runs is shown in Fig. 11. The fact that it resembles Fig. 5 very closely is also a good indication that the equations defining the probabilistic element are operating properly.

Table 8. Kolmogorov-Smirnov Two-Sample Test of the Probabilistic Component in the Watershed Model

| Event Size | Historical Record Cum. Dist. | Run 1 Cum. Dist. | Run 1 $D_n$ (Max. Diff.) % | Run 2 Cum. Dist. | Run 2 $D_n$ (Max. Diff.) % |
|---|---|---|---|---|---|
| > .0002 | .062 | .074 | | .075 | |
| .0003 | .088 | .094 | | .093 | |
| .0005 | .104 | .125 | | .113 | |
| .0007 | .126 | .145 | | .131 | |
| .0010 | .163 | .166 | | .143 | |
| .0016 | .199 | .189 | | .201 | |
| .0020 | .223 | .196 | | .215 | |
| .0025 | .260 | .227 | | .222 | 3.8 |
| .0032 | .287 | .252 | | .253 | |
| .0040 | .298 | .268 | | .269 | |
| .0050 | .305 | .288 | | .280 | |
| .0063 | .311 | .298 | | .308 | |
| .0079 | .329 | .321 | | .328 | |
| .010 | .338 | .332 | | .344 | |
| .016 | .362 | .372 | | .380 | |
| .020 | .393 | .403 | | .400 | |
| .025 | .406 | .423 | | .430 | |
| .032 | .435 | .434 | | .455 | |
| .040 | .466 | .457 | | .466 | |
| .050 | .488 | .495 | | .498 | |
| .063 | .519 | .510 | | .523 | |
| .079 | .543 | .536 | | .552 | |
| .100 | .574 | .559 | | .588 | |
| .16 | .662 | .620 | 4.2 | .647 | |
| .20 | .706 | .666 | | .681 | |
| .25 | .726 | .704 | | .717 | |
| .32 | .757 | .747 | | .747 | |
| .40 | .790 | .783 | | .785 | |
| .50 | .826 | .814 | | .828 | |
| .63 | .863 | .857 | | .860 | |
| .79 | .890 | .883 | | .880 | |
| 1.00 | .912 | .906 | | .900 | |
| 1.60 | .951 | .949 | | .962 | |
| 2.00 | .969 | .959 | | .966 | |
| 2.50 | .976 | .977 | | .982 | |
| 3.20 | .987 | .982 | | .989 | |
| 4.00 | .993 | .990 | | .993 | |
| 6.30 | .996 | .992 | | .995 | |
| 7.90 | | .995 | | | |
| Total No. of Events | 452 | 391 | | 441 | |
| | | $D_{.05}$ = 9.4 | | 9.1 | |

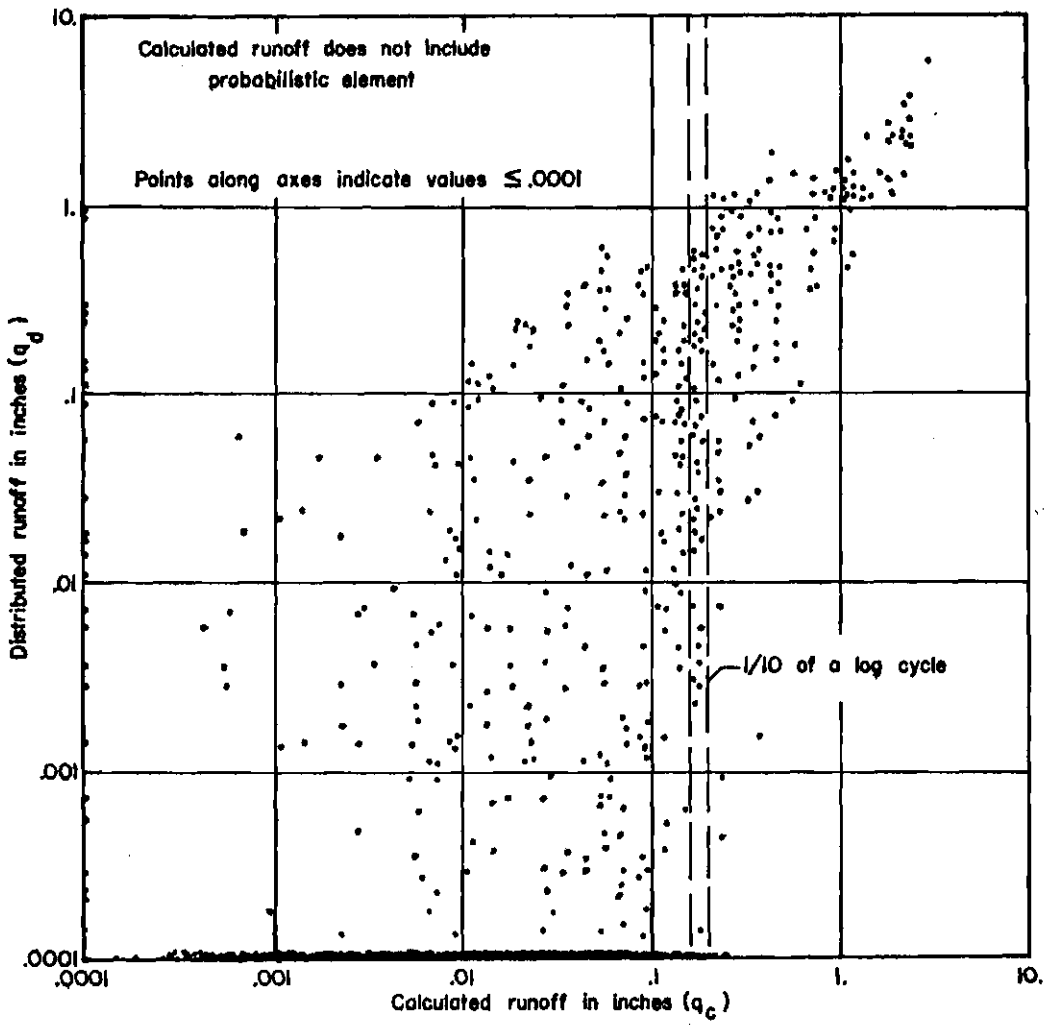



Figure 11. Calculated Runoff vs. Distributed Runoff

# CHAPTER V

## RAINFALL AND EVAPORATION

Inputs to the watershed model are the land use, and the stochastic elements of rainfall, its occurrence and amount, and average daily evaporation. In order that the output from the model be representative of the watershed to which it was fitted, the inputs should have the same statistical properties as the observed record. The 30 years of rainfall and pan evaporation data from 1937-1966 were used as a base from which to develop the generation scheme.

### Rainfall and Evaporation Records

In Chapter IV it was stated that runoff records from the watershed were discontinued during the war years. Many of the rain gages were also discontinued during this period. The U. S. Weather Bureau uses a continuous 30-year period to calculate the average rainfall for a location. This is done to average out wet and dry periods. For this reason, it was desired to have, as nearly as possible, a continuous 30-year period of record upon which to base the generated data. The Blacklands Experimental Watershed had a continuous record of rainfall at the headquarters from December 1937 through 1966, 29-1/2 years. A correlation between this gage, located about 3 miles from the center of the watershed, and the Thiessen weighted rainfall for Station D was used to estimate the Thiessen weighted rainfall for the watershed from July 1, 1943 through March 1, 1949. A linear bivariate regression

between the Thiessen rainfall and the station rainfall yielded

$$P_T = 0.0635 + .8756\ P_S \tag{63}$$

where $P_T$ is the Thiessen weighted rainfall on the watershed and $P_S$ is the rainfall at the headquarters station. Using this equation, it would not be possible to predict any rainfall for the watershed less than 0.06 inch. Neither of the univariate regressions were as good as the bivariate Equation 63, therefore, an equation was developed that went through the means of each group and zero. The equation

$$P_T = 0.997\ P_S \tag{64}$$

was found to be satisfactory and did not distort the record of small events. The standard error of estimate of this equation is about 0.6, but it is a poor measure of the fit to the data because the variances of $P_T$ and $P_S$ are not uniform.

Evaporation data were complete from October 1938 through 1966. The method used to estimate the missing period in 1937-1938 was described in Chapter IV.

## Temporal Distribution of Rainfall

### Using Known Distributions

Several methods may be used to generate a synthetic record of rainfall days. As described in the literature survey, the most commonly used method is to develop, from the historical data, the

distribution of the lengths of wet periods and the distribution of the lengths of dry periods. After fitting a known probability distribution such as the Weibull to these distributions, the days of rainfall are obtained by alternately sampling the two distributions.

Using a Markov Chain

Another method of generating the synthetic record of rainfall days is by using a two-state Markov chain of transition probabilities

$$\begin{array}{cc} & \text{Future state} \\ & \begin{array}{cc} 0 & 1 \end{array} \\ \text{Present state} \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \end{array} \tag{65}$$

where 1 is a wet state, 0 is a dry state, $\alpha$ is the probability of a wet day following a dry day, and $\beta$ is the probability of a dry day following a wet day. This method has been used primarily for short time periods or fractions of a year.

The Markov chain method was selected for this study because the probability of occurrence changes during the year and the transition from one period to the next is much smoother using this method.

Data from the period of record were sorted by month and the transition probabilities, $\alpha$ and $\beta$, calculated for each month from the number and distribution of occurrences of each. The probabilities for each month are shown in Table 9.

Testing the Scheme Used

Ten synthetic 30-year sequences were developed by using random numers and the transition probabilities. To test the adequacy of the

method, two tests were required. A two-tailed t-test at a confidence level of 95 percent was used to test the average number of dry days in the 10 generated samples against the observed number, 1,971 days. The number of dry days, the mean, and the standard deviation of the sequences are shown in Table 10. Since the synthetic sequences were 30 years in length, the observed record was adjusted by direct proportion from 29-1/2 to 30 years, giving 2,034 days. The difference between this value and the mean of the 10 synthetic sequences is 9.1 days. This is considerably smaller than the critical value of 32.56 and the null hypothesis which states "The mean of the synthetic sequences is equal to the observed value" cannot be rejected.

Table 9. Transition Probabilities by Months

| Month | $\alpha$ | $\beta$ | Month | $\alpha$ | $\beta$ |
|---|---|---|---|---|---|
| January | .168 | .656 | July | .098 | .628 |
| February | .206 | .651 | August | .118 | .700 |
| March | .182 | .759 | September | .131 | .622 |
| April | .195 | .704 | October | .122 | .694 |
| May | .208 | .718 | November | .137 | .628 |
| June | .151 | .660 | December | .153 | .698 |

The second test required to insure adequacy of the generating scheme is a test of the distribution of the synthetic sequences. Two of the sequences, selected at random from the group, were chosen for a Kolmogorov-Smirnov two-sample test (described in Chapter A-2) of the cumulative frequency distribution of the number of days between events, i.e., lengths of dry periods. The two distributions are shown in Table 11. The critical value of $D_n$ is about 4.3 for both sequences and the

maximum difference between the sequences and the observed record is about 3.0. Therefore, the null hypothesis which states "The distribution of the synthetic sequence is the same as that of the observed record" cannot be rejected.

Table 10. Number of Dry Days in 10 Synthetic 30-year Sequences

| Sequence | No. Dry Days | | |
|---|---|---|---|
| 1 | 2,077 | | |
| 2 | 2,034 | Mean | = 2043.1 |
| 3 | 2,051 | | |
| 4 | 1,942 | Std. Dev. | = 45.53 |
| 5 | 2,091 | | |
| 6 | 2,059 | Std. Dev. of Mean | = 14.40 |
| 7 | 2,050 | | |
| 8 | 2,015 | $t_{.025}$ 9 | 2.262 |
| 9 | 2,096 | | |
| 10 | 2,011 | Critical Value of U = | 32.56 |

The number of dry days by months of seven additional synthetic series were tested against the observed record using the same criteria as for the yearly data. The average number of events in each month, shown in Table 12, indicates that the synthetic series are from the same populations as the observed records. Since the transition probabilities shown in Table 9 are stable, i.e., show little variation from month to month, and the results of these tests are positive, the Kolmogorov-Smirnov test on individual months was not made.

## Size of the Rainfall Event

In many geographical areas, daily rainfall amounts are serially correlated because of the meteorological characteristics of the area. An equation quite often used to generate data in areas where serial

Table 11. Kolmogorov-Smirnov Two-Sample Test on the Distribution of Lengths of Dry Periods

| Number of Days between Events | Historical Record Cum. Dist. | Sample 1 Cum. Dist. | Sample 1 $D_n$ (Max. Diff.) % | Sample 2 Cum. Dist. | Sample 2 $D_n$ (Max. Diff.) % |
|---|---|---|---|---|---|
| 1 | .324 | .325 | | .327 | |
| 2 | .438 | .421 | | .421 | |
| 3 | .540 | .510 | 3.0 | .511 | 2.9 |
| 4 | .617 | .589 | | .591 | |
| 5 | .671 | .665 | | .662 | |
| 6 | .727 | .717 | | .716 | |
| 7 | .763 | .759 | | .758 | |
| 8 | .799 | .800 | | .801 | |
| 9 | .822 | .829 | | .832 | |
| 10 | .847 | .852 | | .859 | |
| 11 | .874 | .873 | | .880 | |
| 12 | .894 | .893 | | .894 | |
| 13 | .912 | .911 | | .912 | |
| 14 | .924 | .922 | | .929 | |
| 15 | .934 | .931 | | .937 | |
| 16 | .944 | .938 | | .947 | |
| 17 | .951 | .946 | | .954 | |
| 18 | .960 | .955 | | .959 | |
| 19 | .967 | .965 | | .966 | |
| 20 | .970 | .971 | | .972 | |
| 21 | .974 | .973 | | .974 | |
| 22 | .977 | .976 | | .977 | |
| 23 | .980 | .977 | | .982 | |
| 24 | .982 | .982 | | .984 | |
| 25 | .983 | .985 | | .985 | |
| 26 | .988 | .987 | | .987 | |
| 27 | .988 | .990 | | .989 | |
| 28 | .989 | .991 | | .990 | |
| 29 | .989 | .991 | | .994 | |
| 30 | .990 | .991 | | .994 | |
| 32 | .991 | .993 | | .996 | |
| 34 | .993 | .995 | | .997 | |
| 36 | .993 | .999 | | .997 | |
| 38 | .995 | .999 | | .998 | |
| 40 | .996 | .999 | | .998 | |
| Max. Length | 52 Days | 57 Days | | 60 Days | |
| Total No. Events | 1,971 | 2,015 | | 2,039 | |
| $D_{.05}$ (Critical) | | 4.31 | | 4.31 | |

Table 12. Number of Dry Days, by Months, in 7 Synthetic 30-Year Sequences

| Sequence Number | Month | | | | | |
|---|---|---|---|---|---|---|
| | Jan. | Feb. | Mar. | Apr. | May | June |
| 1 | 186 | 195 | 163 | 187 | 211 | 169 |
| 2 | 182 | 217 | 160 | 200 | 166 | 112 |
| 3 | 135 | 201 | 171 | 189 | 236 | 174 |
| 4 | 187 | 206 | 195 | 206 | 222 | 166 |
| 5 | 185 | 200 | 192 | 216 | 231 | 147 |
| 6 | 175 | 193 | 185 | 197 | 206 | 173 |
| 7 | 196 | 201 | 179 | 186 | 204 | 181 |
| Mean | 178 | 201.86 | 177.86 | 194.57 | 215.71 | 168.00 |
| Std. Dev. | 19.97 | 7.92 | 13.74 | 12.53 | 14.06 | 10.65 |
| Std. Dev. of Mean | 7.55 | 3.00 | 5.19 | 4.74 | 5.31 | 4.02 |
| $t_{.05}$ 6 | 2.45 | 2.45 | 2.45 | 2.45 | 2.45 | 2.45 |
| Critical Value of U | 18.49 | 7.34 | 12.72 | 11.60 | 13.02 | 9.86 |
| Obs. Record | 189 | 203 | 180 | 197 | 209 | 168 |
| Difference | 11.00 | 1.14 | 2.14 | 2.43 | 6.71 | 0.00 |
| Accept Null Hyp. | * | * | * | * | * | * |

| Sequence Number | Month | | | | | |
|---|---|---|---|---|---|---|
| | July | Aug. | Sept. | Oct. | Nov. | Dec. |
| 1 | 107 | 152 | 126 | 129 | 172 | 157 |
| 2 | 112 | 124 | 164 | 157 | 164 | 184 |
| 3 | 123 | 114 | 139 | 144 | 161 | 155 |
| 4 | 143 | 154 | 143 | 140 | 164 | 151 |
| 5 | 115 | 147 | 149 | 146 | 183 | 180 |
| 6 | 133 | 112 | 174 | 141 | 163 | 163 |
| 7 | 111 | 128 | 173 | 154 | 160 | 166 |
| Mean | 120.57 | 133.0 | 152.51 | 144.43 | 166.71 | 165.14 |
| Std. Dev. | 13.19 | 17.82 | 18.27 | 9.32 | 8.16 | 12.59 |
| Std. Dev. of Mean | 4.98 | 6.74 | 6.90 | 3.52 | 3.08 | 4.76 |
| $t_{.05}$ 6 | 2.45 | 2.45 | 2.45 | 2.45 | 2.45 | 2.45 |
| Critical Value of U | 12.21 | 16.50 | 16.91 | 8.63 | 7.56 | 11.66 |
| Obs. | 125 | 134 | 156 | 139 | 161 | 173 |
| Difference | 4.43 | 1.00 | 3.43 | 5.43 | 5.71 | 7.86 |
| Accept Null Hyp. | * | * | * | * | * | * |

correlation changes from one period to another, i.e., spring to summer, is

$$P_i = \overline{P}_i + b(P_{i-1} - \overline{P}_{i-1}) + t_i \, s_{p_i} \, (1 - r^2)^{1/2} \qquad (66)$$

in which $P_i$ is the rainfall for day i; $\overline{P}_i$ is the average daily rainfall during the period; $P_{i-1}$ is the rainfall on the previous day; $\overline{P}_{i-1}$ is the average daily rainfall for the period and is equal to $\overline{P}_i$ for long periods of record; b is a regression coefficient; $t_i$ is a standardized, random, normal, and independently distributed variate; $s_{p_i}$ is the standard deviation of $P_i$, and r is the multiple correlation coefficient.

Equation 66 is in theory based on the assumption that the $P_i$ are weakly stationary and normally distributed. As was pointed out in the literature survey, rainfall amounts for short time intervals, such as one day, are not normally distributed. Thus to use Equation 66, they must be normalized by a transform. The inverse transform is then used to change the data generated by the equation back to real values.

Correlation of Rainfall Size with Previous Events

The serial correlation coefficient between rainfall amounts on consecutive days in the period of record was 0.021. At the 95 percent confidence level, a value of 0.080 or higher would be statistically significant for the 639 sets of observations in the sample. The null hypothesis, "rainfall amounts on consecutive days are uncorrelated" was not rejected. The test of the correlation coefficient is identical to the F test for the ratio of the variance accounted for by regression to the error variance.

Analysis of rainfall on days following dry periods gave product-moment correlation coefficients of 0.030 and 0.0023 for the correlation between rainfall amount and number of dry days and previous rainfall, respectively. At the 95 percent confidence level, a value of 0.062 or higher would be statistically significant for the 1,332 sets of observations in the sample. The null hypotheses, "rainfall amounts on days following dry periods are not correlated with the number of dry days or the previous rainfall" were not rejected.

Since these tests indicated that both sets of rainfall, i.e., rainfall following wet days and rainfall following dry days, were essentially independent in amount, the standard t-test for the difference in means was made to see if both sets were from the same population. The mean and standard deviations of the two data sets are shown in Table 13. The calculated z value was 0.289 and the critical value at the 97.5 percent confidence level, since it is a one-tailed test, was 1.96. The null hypothesis, "the means of the two samples are equal" was not rejected.

Table 13. Characteristics of Rainfall Following Wet and Dry Days

| Rainfall Following: | No. of Events | Mean | Standard Deviation |
|---|---|---|---|
| Wet days | 639 | -.009085 | .9617 |
| Dry days | 1,332 | +.004551 | .9527 |

Note: The data used in getting these figures were standardized by month, but no effort was made to separate the events on the basis of antecedent conditions.

A variance ratio or F-test was also made on the two sets of data

to see if the dispersions were the same. The ratio of the two variances was 1.009 and the critical value estimated from the F distribution at a 97.5 percent confidence level was 1.1. Therefore, the null hypothesis, "the standard deviation of the two samples are equal" was not rejected.

As a result of these tests, the rainfall amounts on days following wet days and on days following dry days were assumed to be independent and from the same population. All rainfall events were therefore combined into a single set of data for development of the generating scheme.

The Generation Scheme Selected

The distribution of size of rainfall event by month was examined. No simple normalizing transform of rainfall amounts such as the square root, cube root, and log could be found. However, a skewed log normal transform was satisfactory except for extremely large events.

An equation similar to 66

$$P_{i_j} = \overline{P}_i + t_i^* \, s_{p_i} \tag{67}$$

was fitted to all events on a monthly basis. $P_{i_j}$ is the logarithm of the $j^{th}$ rainfall event in the $i^{th}$ month, $\overline{P}_i$ is the mean of the logarithms of rainfall events in the $i^{th}$ month, $s_{p_i}$ is the standard deviation of the logarithms of rainfall events in the $i^{th}$ month, and $t_i^*$ is a skewed, random, normal, and independently distributed variate as calculated by Equation 56 in Chapter IV.

A test of this system proved to be satisfactory except for extremely large events. Figs. 12 and 13 are cumulative distributions of rainfall events for April and August, respectively. Also shown on the figures are the skewed normal equations. It is obvious from the plots that more events larger than 1.5 to 2 inches will be generated than were recorded in the period of record. These plots are typical of the other months. In only one month, February, did the observed distribution indicate a larger number of events than the skewed normal distribution.

Since extremely large events are an important feature of the study, the distribution from which the data are to be generated was adjusted to conform better with the distribution of observed events. This was accomplished with the aid of a linear transform operating on the skewed random normal variate, t*, when it was larger than a critical value. The linear transform used to adjust the variate is given by

$$t_a^* = t_x^* + T(t^* - t_x^*) \qquad (68)$$

$$t^* \geq t_x^*$$

$$t_a^* = t^*$$

$$t^* \geq t_x^*$$

in which $t_a^*$ is the adjusted variate, $t_x^*$ is the value of the variate at the point where the cumulative probability distributions are equal, i.e., the point where the distributions shown on Figs. 12 and 13 cross

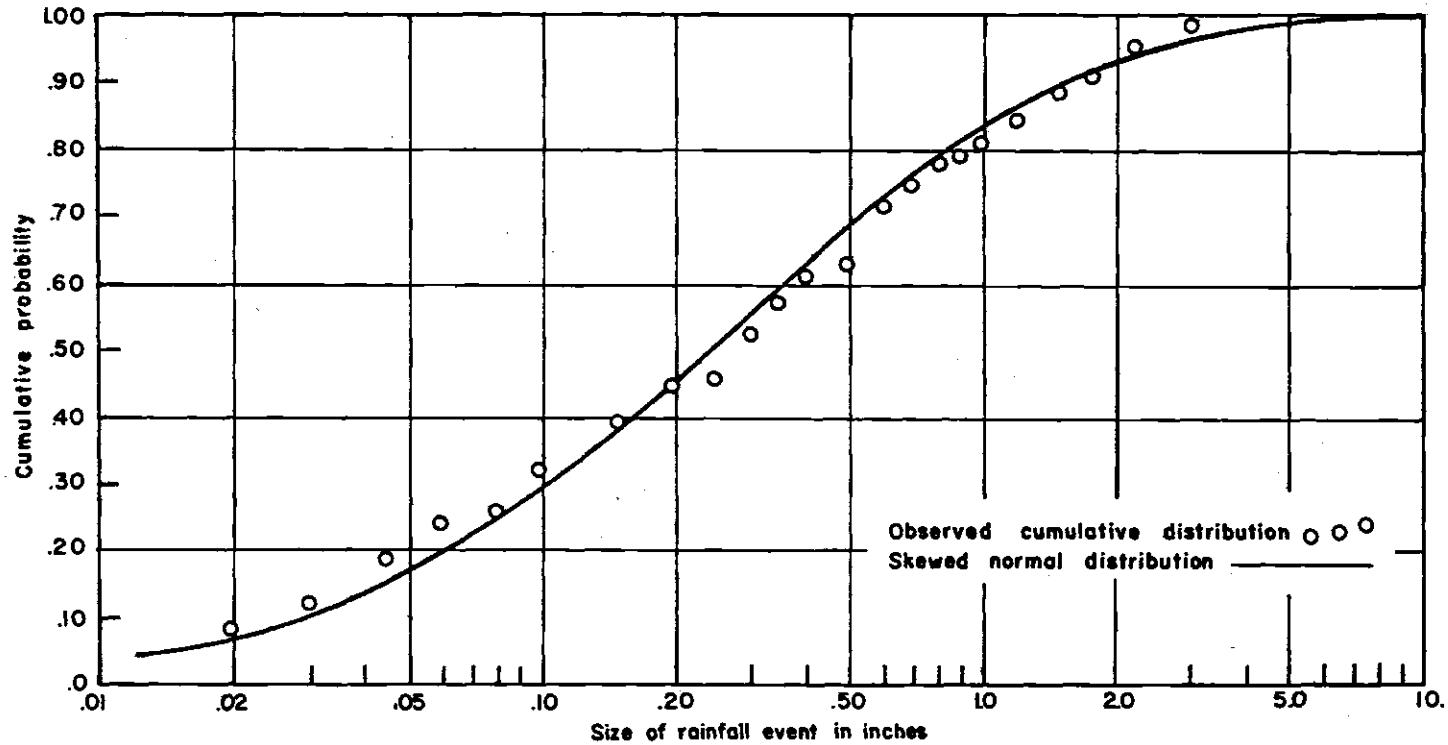



Figure 12. Cumulative Distribution and Skewed Normal Distribution of Rainfall Amounts for April

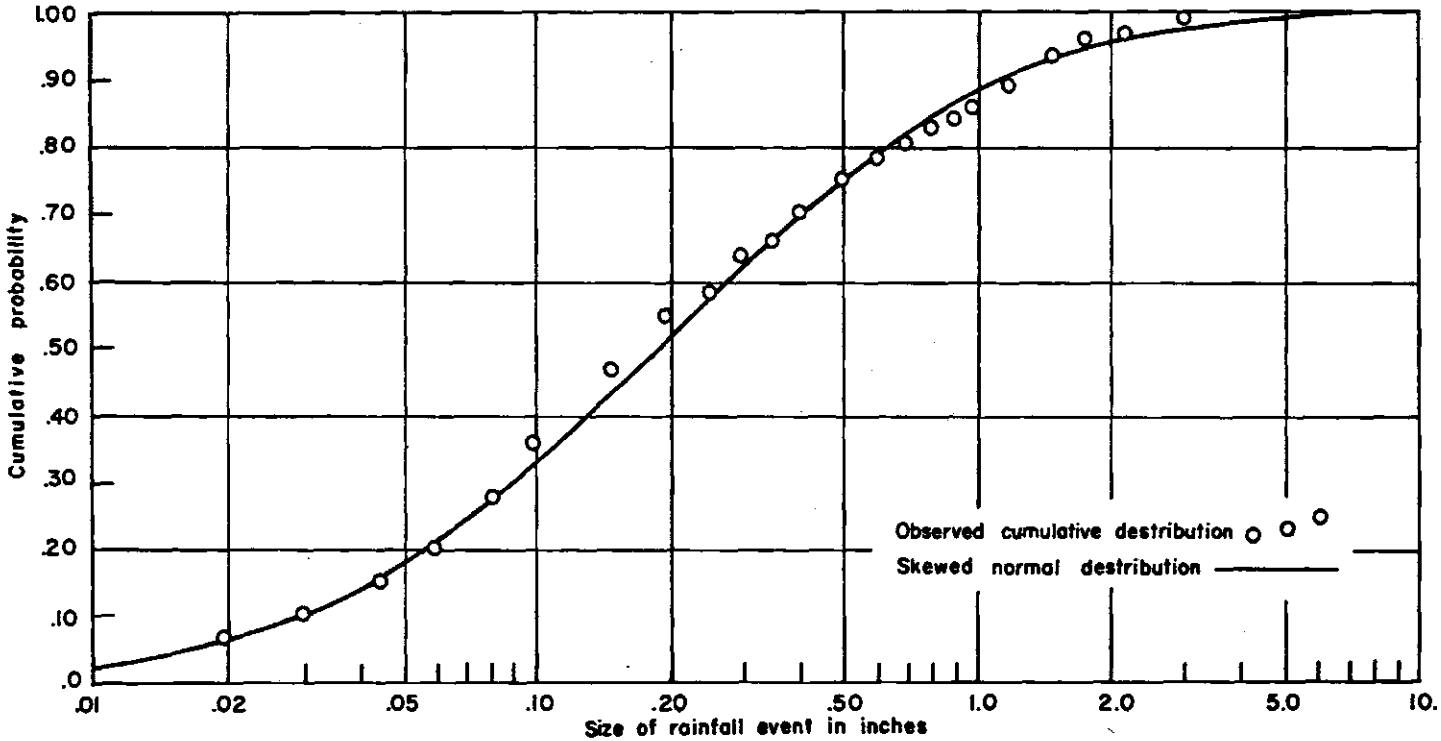



Figure 13. Cumulative Distribution and Skewed Normal Distribution of Rainfall Amounts for August

in the upper right hand corner, t* is the unadjusted variate, and T is a constant. The value of T was calculated such that $t_a$* would be equal to that of the observed distribution at the 98 percent cumulative probability level. The monthly values of $t_x$* and T are listed along with the mean, standard deviation and skew coefficient of the natural logarithms of rainfall events in Table 14.

Table 14. Characteristics of Rainfall Events by Months

| Month | Mean $\overline{P}_i$ | Standard Deviation $s_{p_i}$ | Skew $g_i$ | Critical Value of the variate t* | Transform Coefficient T |
|---|---|---|---|---|---|
| January | -1.911 | 1.503 | -.2189 | .664 | .880 |
| February | -1.727 | 1.411 | -.7378 | 1.095 | 1.19 |
| March | -1.858 | 1.462 | -.2322 | 1.562 | .408 |
| April | -1.486 | 1.560 | -.3254 | 1.414 | .610 |
| May | -1.244 | 1.459 | -.6511 | .932 | .873 |
| June | -1.085 | 1.282 | -.7152 | .772 | .826 |
| July | -1.842 | 1.400 | -.2197 | 1.371 | .649 |
| August | -1.648 | 1.427 | -.0728 | 1.394 | .665 |
| September | -1.820 | 1.714 | -.2999 | 1.123 | .820 |
| October | -1.652 | 1.551 | -.3399 | 1.154 | .800 |
| November | -1.629 | 1.534 | -.2048 | 1.163 | .848 |
| December | -1.661 | 1.484 | -.4491 | 1.068 | .681 |

Note: The mean, standard deviation and skew coefficients are based on natural logarithms of the rainfall data.

Testing the Size Distribution of the Generated Data

Ten synthetic sequences were generated using random numbers, the transition probabilities and the monthly rainfall size distribution. These are the same 10 sequences tested for temporal distribution and presented in Table 10. The number and size of the largest events were very nearly the same as the historical record on each of the months as indicated by a tabulation of the data. Kolmogorov-Smirnov two-sample

tests were made on two synthetic runs for each of the 12 months. The results of these tests for the two months whose distributions are shown in Figs. 12 and 13 are shown in Tables 15 and 16. A summary of the two tests for each of the months showing the maximum difference between the historical and synthetic distributions, $D_n$, the rainfall size interval in which the maximum difference occurs, and the critical value of the maximum difference, $D_{.05}$, are presented in Table 17. The results of all the tests are negative indicating that the null hypothesis, i.e., that the historic and synthetic distributions are the same cannot be rejected.

As an additional check on the generating scheme, the average of the monthly and yearly totals for each of the sequences were compared with the historical record. The two-sided t-test with a confidence level of 95 percent was used to make this test. The results of the test by months, shown in Table 18, indicate that the average monthly and yearly values generated by the system are the same as the historical record, i.e., the null hypothesis could not be rejected on the yearly total and on all months except February and June, and these were borderline cases. This is an excellent test because when one is calculating short time interval events, it is very easy for accumulative errors to cause longer period summaries to deviate from their averages. This can happen even though tests on the adequacy of the scheme may show excellent agreement with observed records.

## Evaporation

Evaporation plays a minor role in the watershed model and only

Table 15. Kolmogorov-Smirnov Two-Sample Test on the Distribution of Size of Rainfall Events for August

| Size of Event | Historical Record Cum. Dist. | Sample 1 Cum. Dist. | Sample 1 $D_n$ (Max. Diff.) % | Sample 2 Cum. Dist. | Sample 2 $D_n$ (Max. Diff.) % |
|---|---|---|---|---|---|
| .01 | .031 | .019 | | .044 | |
| .02 | .063 | .051 | | .062 | |
| .03 | .103 | .077 | | .089 | |
| .04 | .142 | .110 | | .142 | |
| .05 | .182 | .136 | | .178 | |
| .06 | .206 | .162 | | .241 | |
| .08 | .277 | .233 | | .276 | |
| .10 | .349 | .285 | | .339 | |
| .15 | .468 | .363 | | .375 | |
| .20 | .547 | .441 | 10.6 | .464 | |
| .25 | .579 | .513 | | .500 | |
| .30 | .634 | .571 | | .562 | |
| .35 | .658 | .597 | | .562 | |
| .40 | .698 | .636 | | .589 | |
| .45 | .714 | .649 | | .616 | 9.8 |
| .50 | .754 | .675 | | .678 | |
| .60 | .777 | .727 | | .723 | |
| .70 | .801 | .753 | | .758 | |
| .80 | .825 | .772 | | .794 | |
| .90 | .841 | .798 | | .812 | |
| 1.0 | .857 | .811 | | .839 | |
| 1.2 | .888 | .844 | | .875 | |
| 1.4 | .928 | .863 | | .883 | |
| 1.6 | .928 | .889 | | .892 | |
| 2.0 | .960 | .909 | | .946 | |
| 2.6 | .976 | .928 | | .982 | |
| 3.2 | | .961 | | 1.000 | |
| 4.0 | | .974 | | | |
| 5.0 | | .987 | | | |
| 6.5 | | | | | |
| 8.5 | 1.000 | 1.000 | | | |
| No. of Observations | 126 | 154 | | 112 | |
| $D_{.05}$ | | 16.3% | | 17.7% | |

Table 16. Kolmogorov-Smirnov Two-Sample Test on the Distribution of Size of Rainfall Events for April

| Size of Event | Historical Record Cum. Dist. | Sample 1 Cum. Dist. | Sample 1 $D_n$ (Max. Diff.) % | Sample 2 Cum. Dist. | Sample 2 $D_n$ (Max. Diff.) % |
|---|---|---|---|---|---|
| .01 | .036 | .014 | | .020 | |
| .02 | .084 | .053 | | .055 | |
| .03 | .121 | .072 | | .101 | |
| .04 | .163 | .116 | | .137 | |
| .05 | .200 | .135 | | .162 | |
| .06 | .236 | .169 | 6.7 | .192 | |
| .08 | .252 | .213 | | .238 | |
| .10 | .315 | .252 | | .299 | |
| .15 | .389 | .368 | | .411 | |
| .20 | .442 | .436 | | .467 | |
| .25 | .457 | .470 | | .512 | |
| .30 | .521 | .534 | | .583 | |
| .35 | .568 | .577 | | .649 | 8.1 |
| .40 | .605 | .616 | | .665 | |
| .45 | .615 | .635 | | .675 | |
| .50 | .626 | .655 | | .685 | |
| .60 | .710 | .718 | | .720 | |
| .70 | .742 | .752 | | .736 | |
| .80 | .773 | .805 | | .771 | |
| .90 | .789 | .830 | | .807 | |
| 1.0 | .805 | .849 | | .827 | |
| 1.2 | .842 | .878 | | .852 | |
| 1.4 | .884 | .907 | | .878 | |
| 1.6 | .894 | .922 | | .903 | |
| 2.0 | .947 | .922 | | .923 | |
| 2.6 | .973 | .946 | | .954 | |
| 3.2 | .979 | .956 | | .964 | |
| 4.0 | .989 | .971 | | 1.000 | |
| 5.0 | .994 | .980 | | | |
| 6.5 | 1.000 | .985 | | | |
| 8.5 | | .995 | | | |
| No. of Observations | 190 | 206 | | 197 | |
| $D_{.05}$ | | 13.6% | | 13.8% | |

Table 17. Summary by Months of the Kolmogorov-Smirnov Two-Sample Tests on the Distribution of Size of Rainfall

| Month | Sample 1 | | | Sample 2 | | |
|---|---|---|---|---|---|---|
| | Interval of Max. Diff. (in.) | (%) Max. Diff. ($D_n$) | (%) Critical Value ($D_{.05}$) | Interval of Max. Diff. (in.) | (%) Max. Diff. ($D_n$) | (%) Critical Value ($D_{.05}$) |
| Jan. | .04- .05 | 9.94 | 14.10 | .04-.05 | 8.05 | 14.34 |
| Feb. | .04- .05 | 5.87 | 13.57 | .70-.80 | 6.98 | 13.79 |
| Mar. | .06- .07 | 12.07 | 14.16 | .10-.15 | 7.80 | 14.34 |
| Apr. | .05- .06 | 6.69 | 13.68 | .30-.35 | 8.13 | 13.83 |
| May | .35- .40 | 6.01 | 13.22 | .04-.05 | 7.12 | 13.47 |
| June | .09- .10 | 7.10 | 15.02 | .45-.50 | 5.33 | 14.87 |
| July | .07- .08 | 8.11 | 16.69 | .03-.04 | 4.82 | 16.98 |
| Aug. | .15- .20 | 10.60 | 16.33 | .40-.45 | 9.82 | 17.66 |
| Sept. | .03- .04 | 5.07 | 15.87 | .50-.60 | 5.80 | 15.13 |
| Oct. | .15- .20 | 8.02 | 16.44 | .02-.03 | 6.31 | 16.41 |
| Nov. | 1.8 -2.0 | 6.55 | 15.21 | .45-.50 | 4.47 | 15.23 |
| Dec. | .08- .09 | 5.37 | 15.38 | .30-.35 | 3.92 | 15.09 |

Table 18. Average Monthly and Annual Rainfall in 10 Synthetic 30-Year Sequences

| Sequence Number | Jan. | Feb. | Mar. | Apr. | May | June | July | Aug. | Sept. | Oct. | Nov. | Dec. | Year |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.71 | 3.37 | 2.83 | 2.97 | 4.66 | 3.35 | 1.74 | 2.20 | 2.79 | 1.74 | 2.59 | 1.96 | 31.89 |
| 2 | 2.22 | 3.31 | 2.64 | 3.95 | 5.22 | 3.55 | 1.63 | 1.88 | 2.46 | 1.96 | 2.93 | 2.39 | 34.16 |
| 3 | 3.01 | 2.71 | 2.25 | 3.31 | 5.04 | 4.36 | 1.43 | 2.02 | 2.32 | 2.55 | 2.67 | 2.04 | 33.70 |
| 4 | 2.53 | 3.39 | 1.86 | 3.27 | 4.95 | 3.87 | 1.79 | 2.01 | 2.93 | 2.02 | 3.23 | 1.98 | 33.81 |
| 5 | 2.69 | 3.01 | 2.13 | 3.26 | 4.26 | 3.88 | 1.50 | 2.10 | 2.13 | 2.68 | 2.30 | 2.30 | 32.24 |
| 6 | 2.91 | 2.66 | 2.52 | 4.32 | 4.00 | 3.34 | 1.46 | 1.88 | 2.77 | 2.54 | 3.06 | 2.69 | 34.15 |
| 7 | 2.62 | 2.81 | 2.15 | 4.30 | 5.02 | 3.58 | 1.23 | 2.54 | 2.85 | 1.85 | 2.63 | 3.09 | 34.66 |
| 8 | 2.60 | 2.81 | 2.73 | 4.60 | 4.24 | 3.43 | 1.31 | 2.33 | 2.46 | 2.30 | 2.75 | 2.06 | 33.60 |
| 9 | 1.82 | 2.89 | 2.76 | 4.33 | 4.24 | 4.50 | 1.66 | 2.47 | 2.39 | 2.12 | 2.03 | 1.96 | 33.16 |
| 10 | 2.09 | 2.63 | 2.75 | 3.42 | 5.48 | 3.80 | 1.54 | 2.20 | 2.95 | 2.42 | 2.12 | 2.42 | 33.80 |
| Mean | 2.42 | 2.96 | 2.46 | 3.77 | 4.71 | 3.77 | 1.53 | 2.16 | 2.60 | 2.22 | 2.63 | 2.29 | 33.52 |
| Standard Deviation | .44 | .30 | .34 | .59 | .50 | .40 | .18 | .23 | .29 | .33 | .39 | .37 | .87 |
| Std. Dev. of Mean | .14 | .09 | .11 | .19 | .16 | .13 | .06 | .07 | .09 | .10 | .12 | .12 | .23 |
| $t_{.05\ 9}$ | 2.26 | 2.26 | 2.26 | 2.26 | 2.26 | 2.26 | 2.26 | 2.26 | 2.26 | 2.26 | 2.26 | 2.26 | 2.26 |
| Critical Value of U | .32 | .21 | .24 | .42 | .36 | .29 | .13 | .16 | .20 | .23 | .28 | .27 | .62 |
| Observed Record | 2.31 | 2.72 | 2.49 | 3.87 | 4.40 | 3.45 | 1.49 | 2.20 | 2.70 | 2.33 | 2.80 | 2.51 | 33.27 |
| Difference | .11 | .24 | .03 | .10 | .31 | .32 | .04 | .04 | .10 | .12 | .17 | .22 | .25 |
| Accept Null Hypothesis | * | | * | * | * | | * | * | * | * | * | * | * |

the average daily value for the interval between events is needed. For this reason, daily values used to calculate the interval average were calculated directly from generated monthly totals.

Difference in Evaporation Between Wet and Dry Days

It was assumed that there might be a difference in the evaporation on a day of rainfall and that on a dry day. To test this assumption, daily pan evaporation data for the period of record was split into two groups depending upon the occurrence of rainfall. The data were further subdivided by month. A two-tailed t-test for the difference in means of the two groups was then made on a monthly basis. The results of these tests are shown on Table 19. The null hypothesis which states that the means of the two samples are equal was rejected for every month at a confidence level of 95 percent. The critical z value for the test was 1.96.

Since the tests show that evaporation on a wet day is different from that on a dry day, a linear relationship between them based on the difference in means was developed.

$$E_{iw} = C_i E_{id} \tag{69}$$

in which $E_{iw}$ is the average daily evaporation on a wet day in the $i^{th}$ month, $E_{id}$ is the average daily evaporation on a dry day in the $i^{th}$ month, and $C_i$ is the coefficient for the $i^{th}$ month. The coefficients for each month are listed in Table 20.

Using Equation 69, the number of days in the month, the number of and dates of wet days, and the total evaporation for the month, the

Table 19. Tests for the Difference in Evaporation between Wet and Dry Days

| Month | Wet Days | | | Dry Days | | | Diff. in Means | Std. Dev. of the Diff. | Z statistic |
|---|---|---|---|---|---|---|---|---|---|
| | No. of Obs. | Mean $\bar{X}_w$ | Variance $s_w^2$ | No. of Obs. | Mean $\bar{X}_d$ | Variance $s_d^2$ | $X_w - X_d$ | | |
| Jan. | 198 | .0902 | .005117 | 701 | .1100 | .005290 | .0198 | .005778 | 3.43 |
| Feb. | 218 | .0853 | .004662 | 601 | .1291 | .005646 | .0438 | .005479 | 7.99 |
| Mar. | 190 | .1295 | .009166 | 709 | .1694 | .008343 | .0399 | .007746 | 5.15 |
| Apr. | 219 | .1586 | .011853 | 651 | .1854 | .007807 | .0268 | .008131 | 3.30 |
| May | 217 | .1741 | .009832 | 682 | .1961 | .006066 | .0220 | .007362 | 2.99 |
| June | 179 | .2187 | .010407 | 691 | .2414 | .006863 | .0227 | .008251 | 2.75 |
| July | 120 | .2252 | .009339 | 748 | .2784 | .008974 | .0532 | .009477 | 5.61 |
| Aug. | 128 | .2452 | .012870 | 740 | .2851 | .008472 | .0399 | .010583 | 3.77 |
| Sept. | 160 | .2062 | .019150 | 680 | .2361 | .007294 | .0299 | .011420 | 2.62 |
| Oct. | 146 | .1444 | .011315 | 753 | .1941 | .006715 | .0497 | .009296 | 5.34 |
| Nov. | 181 | .1258 | .010471 | 689 | .1573 | .007800 | .0315 | .008317 | 3.79 |
| Dec. | 178 | .0899 | .008546 | 721 | .1241 | .006144 | .0342 | .007519 | 4.59 |

Table 20. Coefficients for Relating Evaporation on Wet Days to that on Dry Days

| Month | Coefficient | Month | Coefficient |
|---|---|---|---|
| January | 0.82 | July | 0.81 |
| February | 0.66 | August | 0.86 |
| March | 0.76 | September | 0.87 |
| April | 0.86 | October | 0.74 |
| May | 0.89 | November | 0.80 |
| June | 0.91 | December | 0.72 |

average daily evaporation between events can be calculated.

The Generation Scheme

In setting up a generating scheme, a check was made to see if the monthly evaporation could be correlated with the month's rainfall or the previous month's evaporation. The correlation coefficients, based on the historic data, were -0.39 and 0.49 for the correlation of present month's evaporation with present month's rainfall and previous month's evaporation respectively. The critical value for the correlation coefficient was 0.135 at a confidence level of 95 percent. The partial correlation coefficients were -0.42 and 0.51 for the correlation of present month's evaporation with present month's rainfall and previous month's evaporation respectively. The critical value of the partial correlation coefficients was 0.110 at the 95 percent confidence level. The null hypotheses which stated that there were no correlations between present month's evaporation and present month's rainfall or between present month's evaporation and past month's evaporation were rejected.

Since both the present month's rainfall and the previous month's evaporation were significantly correlated with the present month's evaporation, a simple Markov model was selected to generate the synthetic monthly evaporation. The equation is:

$$E_i = \overline{E}_i + c_{iE}(E_{i-1} - \overline{E}_{i-1}) + c_{iP}(P_i - \overline{P}_i) + t_i^* s_{Ei}(1 - r^2)^{1/2} \qquad (70)$$

in which $E_i$ and $E_{i-1}$ are the present and past months' evaporation respectively, $P_i$ is the present month's precipitation, $\overline{E}_i$ $\overline{E}_{i-1}$, and

$\overline{P}_i$ are the average monthly evaporation and precipitation, $t_i$* is a skewed, random, normal, and independently distributed variate as calculated by Equation 56 in Chapter IV, $s_{Ei}$ is the standard deviation of $E_i$, r is the multiple correlation coefficient, and $c_{iE}$ and $c_{iP}$ are regression coefficients for the $i^{th}$ month. Values of $c_{iE}$, $c_{iP}$, $s_{Ei}$, r, and the skew coefficient, g, used to calculate $t_i$*, are listed in Table 21 for each of the 12 months.

Table 21. Parameters of the Evaporation Generator

| Month | Coefficient of Previous Month's Evaporation $c_{iE}$ | Coefficient of Present Month's Rainfall $c_{iP}$ | Std. Dev. of Present Month's Evaporation $s_{Ei}$ | Multiple Correlation Coefficient r | Skew Coefficient g |
|---|---|---|---|---|---|
| Jan. | .427 | -.181 | .716 | .490 | .115 |
| Feb. | .028 | -.215 | .812 | .796 | - .338 |
| Mar. | .410 | -.289 | 1.316 | .781 | .505 |
| Apr. | .374 | -.147 | 1.139 | .604 | .795 |
| May | .508 | -.155 | 1.257 | .632 | - .051 |
| June | .377 | -.183 | .973 | .622 | .348 |
| July | .899 | -.352 | 1.719 | .687 | .012 |
| Aug. | .799 | -.224 | 1.612 | .909 | .729 |
| Sept. | .469 | -.340 | 1.408 | .837 | .293 |
| Oct. | .476 | -.174 | 1.305 | .843 | .879 |
| Nov. | .405 | -.160 | .985 | .664 | .028 |
| Dec. | .451 | -.101 | 1.070 | .870 | 1.010 |

Testing the Evaporation Generator

Equations 70, 68, 67, and 65 were combined in a computer program and used with the random number generator (described in Chapter A-1) to develop a 30-year sequence of synthetic monthly pan evaporation. Statistical t and F tests were used to see if the mean and variance of the generated data were different from the historic record. The results,

by months, are shown in Table 22. The significance levels for both the two-tailed t-test and the F test were set at 95 percent. These gave critical values of 1.96 and 1.84 for the difference-in-means test and variance-ratio test, respectively. The null hypotheses which stated that the means and variances of the two sets of data were equal could not be rejected for any month.

Table 22. Tests for the Difference between Generated and Observed Evaporation Data

| Month | Historic n = 28 | | Generated n = 30 | | Diff. in Means | Std.Dev. of the Diff. | z Statistic | F Variance Ratio |
|---|---|---|---|---|---|---|---|---|
| | $\bar{E}_H$ | $s_{E_H}$ | $\bar{E}_G$ | $s_{E_G}$ | $\bar{E}_H - \bar{E}_G$ | | | |
| Jan. | 3.231 | .692 | 3.279 | .591 | .048 | .170 | .282 | 1.368 |
| Feb. | 3.212 | .586 | 3.372 | .665 | .160 | .164 | .971 | 1.290 |
| Mar. | 4.836 | 1.052 | 4.703 | .944 | .133 | .263 | .504 | 1.243 |
| Apr. | 5.305 | 1.122 | 4.801 | .828 | .504 | .260 | 1.934 | 1.837 |
| May | 5.848 | 1.209 | 5.483 | 1.237 | .365 | .321 | 1.134 | 1.047 |
| June | 7.067 | .972 | 7.094 | .993 | .027 | .262 | .100 | 1.042 |
| July | 8.394 | 1.719 | 8.438 | 1.313 | .044 | .404 | .110 | 1.714 |
| Aug. | 8.656 | 1.612 | 8.699 | 1.277 | .043 | .350 | .123 | 1.595 |
| Sept. | 6.912 | 1.408 | 7.188 | 1.311 | .276 | .358 | .771 | 1.153 |
| Oct. | 5.674 | 1.120 | 5.691 | .903 | .017 | .268 | .060 | 1.539 |
| Nov. | 4.441 | .901 | 4.400 | 1.132 | .041 | .268 | .154 | 1.581 |
| Dec. | 3.477 | .708 | 3.471 | .666 | .006 | .181 | .030 | 1.129 |

Since all the tests made thus far on the scheme for generating rainfall and runoff have shown the system to be computationally sound, the system was assumed to be operational.

# CHAPTER VI

# DESIGN OF EXPERIMENT

## Introduction

In Chapter I four objectives of the research were outlined. The first and primary objective was to demonstrate the use of multiple discriminant analysis in the study of hydrologic data. To demonstrate its use, the problem of recognizing the differences in hydrologic records from a modeled watershed with two or more land use patterns was selected. For such an analysis various hydrologic indices were used as discriminators of land use change. By using the techniques described in Chapter III, an optimum number of discriminators or indices are selected. Linear combinations of these are used to calculate discriminant scores which were in turn used to classify the individual observations into the group to which they probably belong.

The second objective was to determine the effect that the degree of land use change has on the ability to distinguish hydrologic differences in a modeled watershed. To evaluate this objective, several different land use patterns having different degrees of change were superimposed on the watershed model. The results were evaluated by multiple discriminant analysis.

The third objective was to determine how long the period of record would need to be in order to be able to distinguish hydrologic differences as a result of changes in land use of a modeled watershed.

This is, of course, a function of the degree of land use change. To evaluate this objective, the effects of the different land use changes superimposed on the model were analyzed at five different periods or lengths of record.

The last objective which was to develop a technique for generating synthetic climatic data was necessary in order to be able to evaluate the second and third objectives. This objective was met by combining the computer program which generates rainfall and evaporation with the watershed model. Using these computer programs together, a continuous record of synthetic rainfall, evaporation, soil moisture, and runoff are obtained. Land use can be changed at any time without interrupting the cycle and periods of record of any length can be obtained.

## Synthetic Data Sets

The watershed model described in the previous chapters takes synthetically generated rainfall and evaporation and uses these to calculate volumes of daily (storm) runoff. Data used in the discriminant analyses to represent a given land use pattern are developed from the daily runoff volumes. One observation or data point for the discriminant analysis is made up of 64 hydrologic characteristics taken from the synthetic data during a given period of record. The initial discriminant analyses were made on 30-year summaries or periods of record because it is generally accepted, as stated in Chapter V, that this is a long enough period to average out wet and dry cycles.

Data sets for each of the different land use patterns investigated were limited to 50 observations each because of the time required to generate this quantity of data, i.e., 1,500 years per set of 50 observations. The random number generator used in the system, described in Appendix Chapter A-1, had a recycle period of approximately one-half billion numbers; therefore no two "observations" were the same.

The 64 hydrologic characteristics were summarized from the generated data at 2-, 5-, 10-, and 20-year intervals in addition to the summary at the 30-year interval. The 64 characteristics are:

(1) Rainfall events greater than 5 inches.

(2) Total number of rainfall and runoff events.

(3) Average and standard deviation of runoff, soil moisture, initial abstraction, and precipitation.

(4) The number of rainfall events, the number of runoff events, the percent of rainfall events producing runoff, and the average and standard deviation of runoff events in each of the following ranges of rainfall:

| | |
|---|---|
| 0 - 0.5 inch | 2.0 - 2.25 inches |
| 0.5 - 0.75 " | 2.25 - 2.50 " |
| 0.75 - 1.00 " | 2.50 - 2.75 " |
| 1.00 - 1.25 inches | 2.75 - 3.0 " |
| 1.25 - 1.50 " | 3.0 - 3.5 " |
| 1.50 - 1.75 " | 3.5 - 4.0 " |
| 1.75 - 2.0 " | 4.0 - 5.0 " |
| | > 5.0 " |

(5) Annual maximum runoff events.

(6) Average, standard deviation and skew of the logarithms of the annual series for summary periods of 10 years or more.

(7) Average number of days between runoff events equal to or greater than the following:

| | |
|---|---|
| 0.1 inch | 1.0 inch |
| 0.2 " | 2.0 inches |
| 0.4 " | 5.0 " |
| 0.7 " | |

Most of these data can be used as discriminators or indices in the discriminant analyses. However, such items as the total number of rainfall events and the mean and standard deviation of soil moisture, initial abstraction, and precipitation are either not related to the land use or not practical to use. The remaining items, 58 in number, were the indices of hydrologic change used in the discriminant analyses described in the following chapters.

## Land Use Patterns

The land use patterns selected for use in the study were based on the historic record presented in Table 1, Chapter IV. A brief examination of this table shows that three land uses were predominant at different times. These were cultivated row crops in 1937, native grass meadow in 1961, and Bermuda pasture in 1966. Cultivated oats and no crops were of minor importance compared to the others. The land use combinations for these three years were selected as representative of fairly extreme land use conditions and were therefore used as the patterns for the first three groups studied.

Results of discriminant analyses and a plotting of the data in discriminant space showed that Group III, Bermuda pasture predominant, was isolated somewhat from the other groups. Since Group I, cultivated row crops predominant, and Group II, native grass meadow predominant, were quire similar, only one, cultivated row crops, was selected for comparison with Bermuda pasture in studying how magnitude and extent of land use change affect the ability to differentiate between land use conditions. Land use for the first three groups studied and the four additional groups examined are presented in Table 23.

Table 23. Land Use Patterns Used in the Synthetic Data Sequences as a Percent of the Total Area

| Group | Cultivated Row Crops | Native Grass Meadow | Bermuda Pasture | Cultivated Oats | Cultivated No Crops |
|---|---|---|---|---|---|
| I | 72 | 9 | 11 | 1 | 7 |
| II | 21 | 49 | 20 | 3 | 7 |
| III | 15 | 8 | 69 | 6 | 2 |
| IV | 50 | 8 | 34 | 4 | 4 |
| V | 34 | 8 | 50 | 4 | 4 |
| VI | 32 | 56 | 4 | 4 | 4 |
| VII | 4 | 56 | 32 | 4 | 4 |

The last four groups were set up to compare Bermuda pasture and cultivated row crops only. Additional groups required to isolate other land use interrelationships would cost more than it was felt could be justified by the study. A detailed discussion of the plot, the group separation, and the reasons for selecting Bermuda pasture and cultivated row crops is contained in the Appendix, Chapter A-III.

Two things were considered in selecting land use patterns for Groups IV through VII in order to best complement the analyses of

Groups I and III: (1) Which is more important on a modeled watershed - a nearly complete change in land use on a small part of the watershed, or a partial change on a large part of the watershed, and (2) how much of a change in land use is necessary in order to show a significant difference in modeled hydrologic characteristics? Groups IV and V, shown on Table 23, represent a watershed in which a partial change from one land use to another took place on a large part of the watershed, i.e., 84 percent of the watershed was in the two major land uses. However, the amount of actual change was on only 16 percent. Groups VI and VII, also shown on Table 23, represent a watershed in which a nearly complete change in land use took place on a small part of the watershed, i.e., 36 percent of the watershed was in the two crops. Yet the amount of actual change was 28 percent. In contrast to both of these pairs, Groups I and III represent a watershed in which a nearly complete change in land use took place on a large part of the watershed. The amount of actual change was about 56 percent.

## Length of Record

It was mentioned in the introduction that five different summary periods were used to find how the period of record relates to the ability to distinguish changes in land use. Climatic summaries were collected at 2-, 5-, 10-, 20-, and 30-year intervals. The intervals were selected on the assumption that there would be good discrimination at the 30-year summary period and no discrimination at the 2-year summary period, and that the transition from ability to discriminate to the lack of ability to discriminate would take place somewhere between

these two extremes. The initial analyses which were used to study the first three groups as described in the previous section were made on the data from the 30-year summaries and showed good discrimination. The same three groups were then analyzed using the data from the 2-year summary period. Results showed that although the discrimination was poor, it was statistically significant at the 95 percent confidence level. As a result, the data for the 10-year summary period was analyzed but the 5- and 20-year periods were not. It was not felt that enough information could be gained to make it worth the time and cost necessary for the analyses.

# CHAPTER VII

## DISCUSSION OF RESULTS

In Chapter III of the Appendix, the results of applying multiple discriminant analysis to sets of synthetically generated hydrologic data are presented. The synthetic data were generated to help answer two important questions of interest to hydrologists and others involved in the design of water resources systems and to demonstrate the use of multiple discriminant analysis:

(1) What effect does length of record have on the ability to distinguish hydrologic differences in a modeled watershed?

(2) What effect does degree of land use change have on the ability to distinguish hydrologic differences in a modeled watershed?

Answers to these questions were based on synthetic data rather than actual data because it is not often that one can find hydrologic records of a watershed under more than one land use in which the period of record under each is long enough to be free of climatic effects. A second and possibly more important reason for using synthetic data is that the hydrologic design of many structures is based on data from a watershed model rather than real watershed data.

### Use of Multiple Discriminant Analysis

The primary objective of this study was to demonstrate the applicability of multiple discriminant analysis to the study of hydrologic data. In Chapter III the mathematics of multiple discriminant

analysis are presented along with a review of multivariate statistical tests applicable to discriminant analysis. In the third chapter of the Appendix, several groups of hydrologic data are subjected to multiple discriminant analysis to see if the individual groups can be identified and if so, how significant the group differences are. In the next two sections of this chapter the results of applying discriminant analysis to hydrologic data are presented.

Selecting Significant Variables

Wallis (37) showed that by using a dummy variable, he could very efficiently use principal components analysis and varimax rotation to select the most sifnificant variables from the original set of variables for use in discriminant analysis. Principal components analysis is a statistical technique used to obtain the orthogonal components of a matrix and varimax rotation is a technique for maximizing the variance of a matrix in order to isolate its strongest elements. Of five methods of selecting variables, Wallis found that two methods, both based on components analysis and varimax rotation, were best. Of the two, the reduced rank method produced consistently better results on smaller sample sizes. Therefore, it was used to select primary variables for use in this study. A stepwise selection of variables as described in Chapter VI was then performed on the primary variables to get the statistically significant variables. It was found that, for the statistically significant variables, the order of selecting variables by the stepwise procedure was identical to the order obtained from the reduced rank method. The criterion used in the reduced rank method for determining the order was the size of factor loading on the dummy

variable, i.e., if the factors were ranked according to the loading on the dummy variable, then the variables selected by the reduced rank method from these will be in order according to significance. Cooley and Lohnes (6) state that the size of the elements in the scaled discriminant vector are an indication of the variables contribution to discrimination. It was found, in the cases where discrimination was good, that the scaled vector loadings were in agreement with the order of selection of variables.

Thus, all three methods show the same ranking of variables. However, neither the reduced rank method nor the scaled vector approach indicate how many of these variables are significant. The stepwise selection of variables appears to do both; select the order and the number of variables. The stepwise selection of variables however is very inefficient except for a small number of variables because it must search through all variables each time it adds a variable to the list. The combination of using components analysis to rank the variables and stepwise selection to select the number is the most efficient from the standpoint of computer time. Mahalanobis' $D^2$, which is a measure of group difference, was the criteria used to determine the significance of added variables in the stepwise procedure. The difference in $D^2$ scores, $\Delta D^2$, based on the addition of one variable is distributed as $\chi^2$ and was used to test the significance of the variable. The critical $\chi^2$ value used in the test is a function of the number of predictors, groups, and variables selected as described in Chapter III. An independent Brier and Allen test based on the predictive ability of an equation based on two variables over and above one based on one

variable showed that the $\Delta D^2$ was a satisfactory indicator of the cutoff point in the stepwise selection of variables. (See the two-year summary period described in Chapter III of the Appendix.)

Although the concept of a dummy variable for indicating group membership cannot be extended beyond the two-group case, the technique was used in a composite form for reducing the number of variables in the three-group case. It was accomplished by taking the three groups, two at a time, and using the dummy variable to select the significant variables. The three sets were then combined for the discriminant analysis.

Selecting Significant Discriminant Functions

The number of linear discriminant functions that can be developed from a data set is either P or G-1 where P is the number of predictors and G is the number of groups, whichever is the smallest. If not all of the initial P predictors are selected, then the number of functions is either R or G-1 where R is the number of predictors selected. The significance of these functions is a function of the size of the latent roots of the $W^{-1}B$ matrix and the number of groups and predictors. Transformations of these roots are distributed as $\chi^2$ thus their significance can be tested. Two independent Brier and Allen tests based upon prediction with one and two discriminant functions substantiated results based on the $\chi^2$ test. In both cases, described in Chapter III of the Appendix, only one root was statistically significant. The critical $\chi^2$ value used in the test is a function of the number of predictors, groups, and number of roots selected as described in Chapter III. The significance level used in all of the tests was 0.05.

Group Classification

The discriminant functions for the various groups were used to calculate discriminant scores for each of the observations. These in turn were used to assign the observation into one or the other of the groups. Tables were prepared which showed how the assigned groupings compared to the observed groupings. Plots were also prepared of the distribution of discriminant scores. Both the tables and plots are used to give a visual representation of the degree of discrimination.

Significance of Discrimination

Wilks' $\Lambda$ and Box's M were used to test the hypotheses; $H_2$, equality of means, and $H_1$, equality of dispersions, respectively. In all cases the test of $H_2$ showed significant differences in the centroids, i.e., in discrimination ability. In all but two cases, the $H_1$ test showed that the dispersions were equal. The reliability of the test of $H_2$ is dependent upon equality of group dispersion. However, Cooley and Lohnes (6) have stated that the test of $H_2$ is relatively insensitive to moderate departures from homogeneity of dispersion. Figs. 14 and 18, discussed in more detail later, show the distribution of points for the two cases where the dispersions are different. They still show good discrimination and would substantiate the claim by Cooley and Lohnes. Thus the two cases where the dispersions were found to be different may still represent a significant degree of discrimination.

Summary

In summary it was found that multiple discriminant analysis when combined with component analysis and varimax rotation, provides a

powerful tool for evaluating differences between groups. In the two-group case, components analysis and varimax rotation, used in conjunction with a dummy variable, can be used to find the variables most indicative of the difference between the groups. The significant variables in this list can then be found by adding the variables in their order of significance and testing $\Delta D^2$ with $\chi^2$ tables. In the three-group case, the number of variables can be reduced by using components analysis and varimax rotation in conjunction with a dummy variable by taking the groups two at a time. The most significant variables from these three pairs can then be subjected to a stepwise selection of variables using the $\Delta D^2$ criteria as a cutoff point. The number of significant roots or discriminant functions is found by use of the $\chi^2$ test on the eigenvalues of the $W^{-1}B$ matrix. Wilks' $\Lambda$ and Box's M are used to test the significance of discrimination and equality of dispersions respectively. Results of classification and a plot of the group dispersions show the degree of discrimination.

## Effect of Length of Record on Land Use Discrimination

In this section of the discussion the effect of the length of the synthetic sequences on the output of the watershed model is examined. It is in response to the third objective of the research.

Multiple discriminant analyses of Groups I, II, and III were made from data summaries equivalent to lengths of record of 2, 10, and 30 years. Details of these analyses are presented in reverse order in the third chapter of the Appendix. In all three cases it was found that the number of discriminant functions that would contribute significantly

to the discrimination was only one. The second root was not statistically significant. It was also found that only a few variables were needed to perform the discrimination; 1 in the 2-year summary period, 4 in the 10-year summary period, and 6 in the 30-year summary period. The variables found to be statistically significant are shown in Table 24. The numbers in the last three columns show the order of significance of the variables. Variable notation is described in Chapter A-III in the Appendix.

Table 24. Variables Used in the Discriminant Analysis of the 2-, 10-, and 30-Year Summaries

| No. | Variable Identification | 2-Year Summary | 10-Year Summary | 30-Year Summary |
|---|---|---|---|---|
| 2 | $\overline{RO}$ | | 2 | 2 |
| 13 | r-1 %RO | | 3 | 1 |
| 15 | r-1 $\overline{RO}$ | | | 3 |
| 20 | r-2 %RO | 1 | | |
| 22 | r-2 $\overline{RO}$ | | | 5 |
| 27 | r-3 %RO | | 1 | |
| 65 | r-8 $s_{RO}$ | | 4 | |
| 76 | r-10 %RO | | | 4 |
| 117 | $AS_g$ | | | 6 |

$\overline{RO}$ is the average runoff for the period
r-i %RO is the percent of precipitation events in range i that produce runoff
r-i $\overline{RO}$ is the average runoff event from range i
r-i $s_{RO}$ is the standard deviation of runoff events in range i
$AS_g$ is the skew coefficient of the annual series of maximum events.

Most discrimination comes from runoff characteristics of the lowest three ranges of rainfall; that is from rainfall events less than one inch. This is an indication that simulated runoff from the larger rainfall events is more a function of the rainfall than of the land use; whereas, runoff from small rainfall events is more a function of the

land use parameters. These findings are logical, however it would not be possible to deduce this from an inspection of the watershed model because of the interaction between soil moisture, land use, the climatic variables, and the probabilistic element. The table also shows that the primary discriminators for the shorter lengths of record are the percent of rainfall events producing runoff whereas in the 30-year period of record, the average amount of runoff per storm is also important. The average runoff per storm is probably not as important in the shorter lengths of record because the variance of this variable is large and this would tend to mask out any significant difference due to land use change.

Tests on the overall discriminant ability of the variables showed that there was a significant difference at the 95 percent confidence level between the simulated hydrologic characteristics from the watershed under different land uses at all three lengths of record. However, judging from the size of the calculated F value, the ability to discriminate between the groups was considerably less at the 2-year level than at the 30-year level. This is also substantiated by the results of the classification program which show 73 percent correctly classified observations at the 30-year level, 57 percent at the 10-year level, and 47 percent at the 2-year level.

To get a better feel for the difference between groups at the three different summary levels, the distribution of discriminant scores for the three groups at the three summary levels, 30, 10, and 2 years, were plotted. The three plots are respectively, Figs. 14, 15, and 16. Each of the distributions was assumed to be normal. This assumption

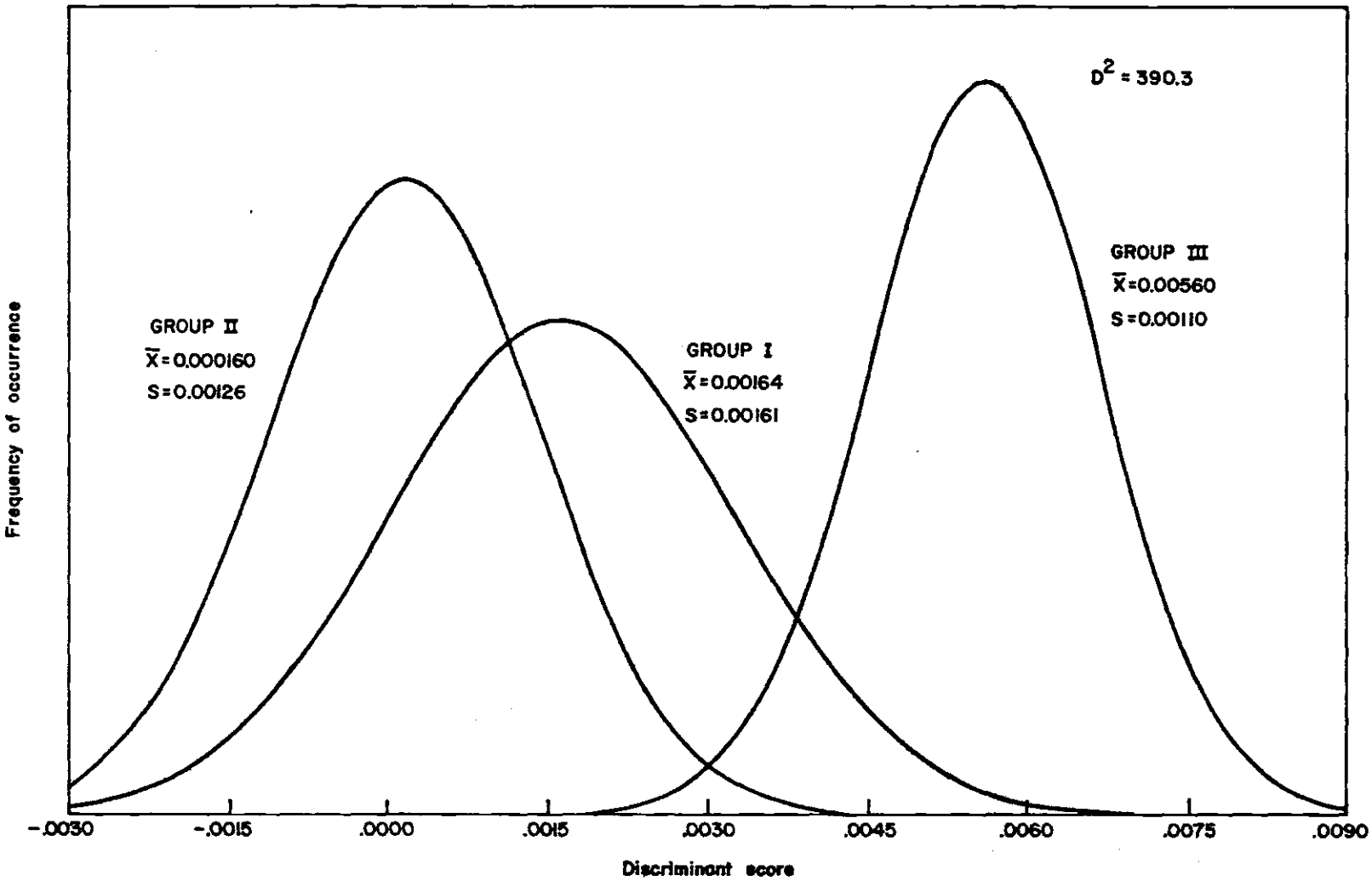



Figure 14. Distribution of Discriminant Scores for Groups I, II, and III for the 30-Year Summary Level

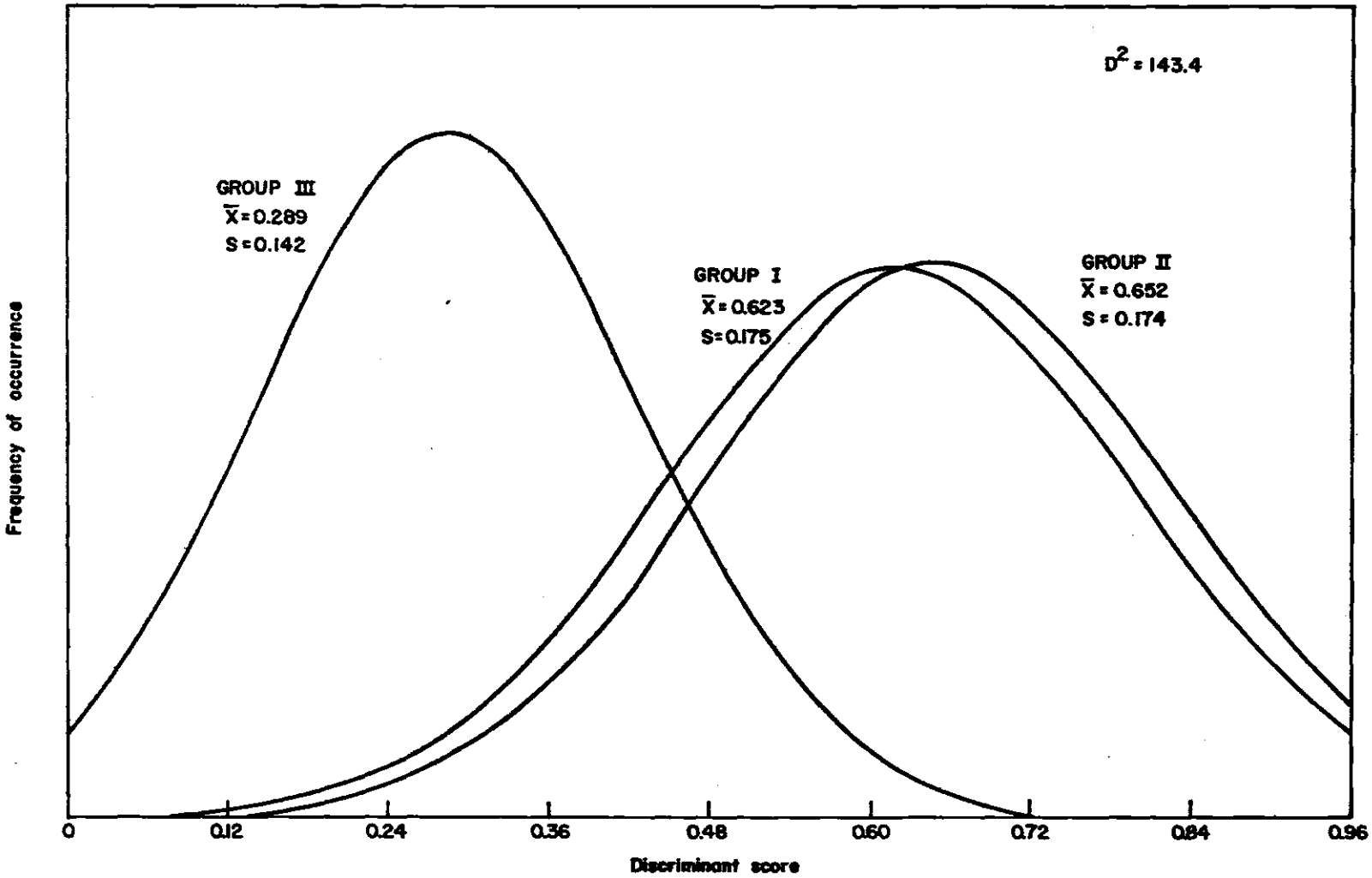



Figure 15. Distribution of Discriminant Scores for Groups I, II, and III for the 10-Year Summary Level

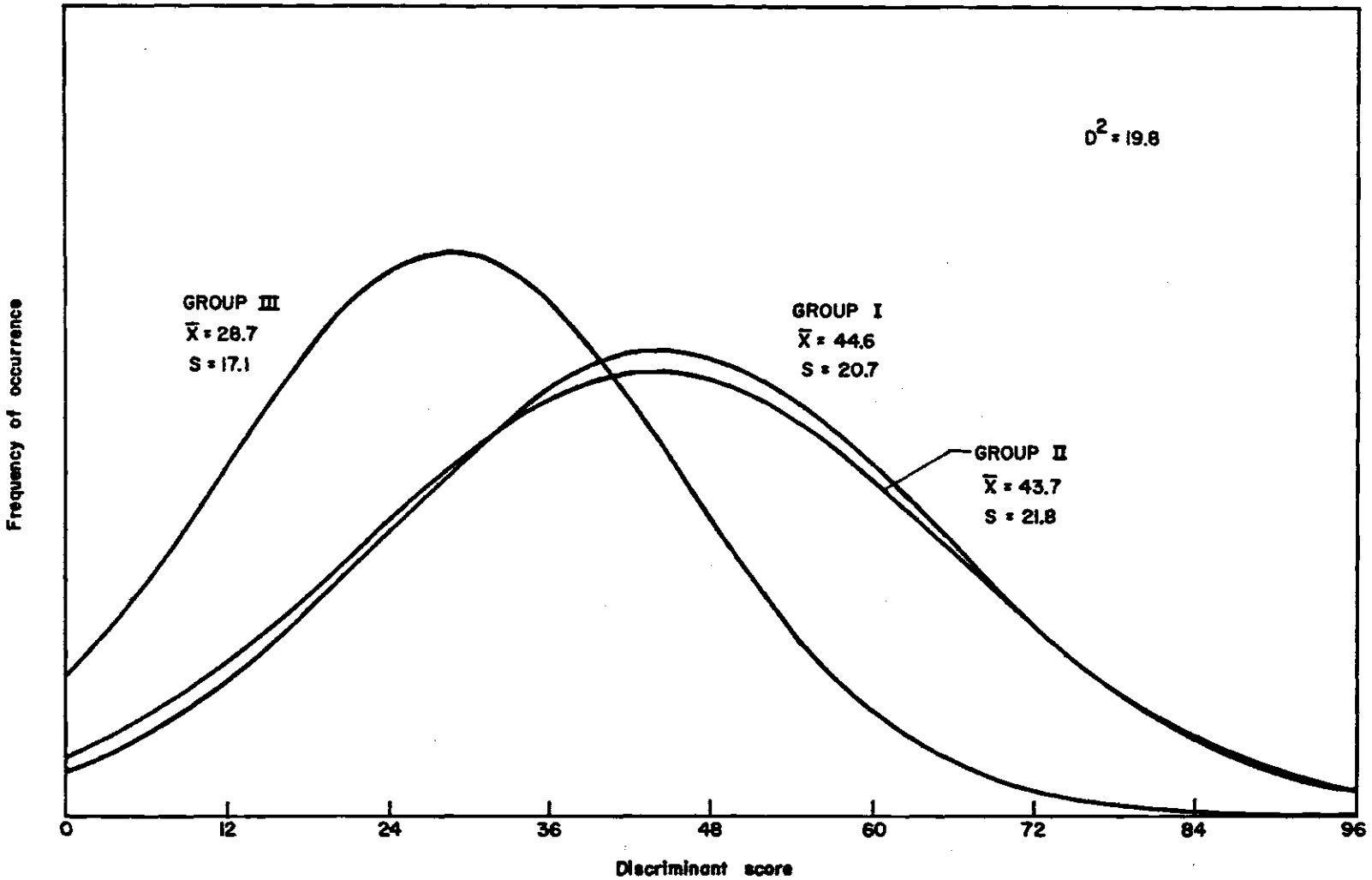



Figure 16. Distribution of Discriminant Scores for Groups I, II, and III for the 2-Year Summary Level

was substantiated by use of the $\chi^2$ test; the results of which are presented in the discussion of Groups IV and V. The scales on the X axes of the three figures cannot be compared with one another because they develop from a normalizing of the discriminant function coefficients. However, the three drawings were scaled such that the dispersions of the groups relative to the overall spread could be seen. It is quite obvious from the figures that the ability to discriminate between the three groups becomes less as the length of the summary period becomes less. It is also obvious from the drawings that the group dispersions are not equal for the 30-year summary. This is in agreement with the test of $H_1$, assumption of equality of dispersions, which was rejected for this case. For both of the other groups the size of the F value for the test of $H_1$ and the drawings show that the test is not rigorous although it was significant at the 95 percent confidence level. The three drawings also indicate that Group III, Bermuda pasture, is quite different from the other two groups; cultivated row crops, Group I, and native grass meadow, Group II. In fact, Groups I and II are almost indistinguishable at the 2-year and 10-year summary levels.

Figure A4 gives a better picture of how the three groups are separated. It is a bivariate plot of the two discriminant scores and is described in the third chapter of the Appendix.

The reason why Bermuda pasture seems to be so much different from the other two land uses has a physical explanation. Bermuda pasture is used in the Blacklands area as a stomp lot for cattle and as such is highly overgrazed. It is also tramped by the cattle, thus reducing the infiltration rate of the soil. Both of these things will

tend to produce a higher rate of runoff from the larger storms than would be produced on either cultivated row crops or native grass meadow. This is shown by looking at the average runoff for the three groups shown on Table A14, variable number 2. It is interesting to note that although the average runoff is greater, the percent and average runoff from small events is less than that on the other two groups (see variables 13, 15, and 22 on the same table). A higher percent of rainfall events producing runoff is noted for events of about 2.5 inches; variable 76. This may be explained by the fact that the combined action of tramping and overgrazing could provide a slightly higher initial abstraction in the form of increased surface detention thus reducing the runoff from small events. However, the lower infiltration rate will produce more runoff from larger storms.

It appears from these results that differences in hydrologic characteristics of a watershed, first under one land use then under another, may not be distinguishable if short periods of record are used in the analysis. The tests described above show that it is possible to distinguish differences in sets of observations at all three lengths of record. However, the variance of the means which is really what has been evaluated is proportional to the square root of the number of events; thus, if only one set of observations is available for each land use pattern, it is very likely possible that differences would not be statistically significant.

## Effect of Degree of Change on Land Use Discrimination

In this section of the discussion, the effect that the degree of

land use change has on the ability to distinguish hydrologic differences in a modeled watershed is examined. It is in response to the second objective of the research.

Three different levels of land use change were evaluated with the multiple discriminant analysis program. The results of these analyses are presented in the third chapter of the Appendix. The study compares the separation of Groups I and III with the separation of Groups IV and V and the separation of Groups VI and VII. In Groups I and III the two land uses under consideration are Bermuda pasture and cultivated row crops. Combined, they represent about 84 percent of the watershed with about 70 percent representing one or the other of the two land uses; thus the change from Group I to Group III is a nearly complete change in land use on a large part of the watershed. It also represents an absolute change from one land use to another of about 56 percent.

In Groups IV and V the two land uses under consideration are the same as those for Groups I and III; they also represent 84 percent of the watershed. In both of these groups, 50 percent of the watershed is in one or the other of the two land uses, therefore the change from Group IV to Group V is a partial change of land use on a large part of the watershed. It also represents an absolute change of only 16 percent.

In Groups VI and VII the two land uses under consideration are the same as those for Groups I and III; however, they represent only 36 percent or about one-third of the watershed. In these groups, 32 percent of the watershed was in one or the other of the two land uses so that a change from Group VI to Group VII is nearly a complete change

in land use on a small part of the watershed. It represents, however, an absolute change from one land use to another of about 28 percent.

The distribution of all land uses for the six groups was presented in Table 23.

Multiple discriminant analyses were performed on these three pairs of data. The results and details are presented in the third chapter of the Appendix. Since only two groups were compared at one time, only one discriminant function existed. As described for the three-group cases, the number of variables needed to differentiate between the groups was very small. The variables found to be statistically significant are shown in Table 25. The numbers in the last three columns show the order of significance of the variables.

Table 25. Variables used in the Discriminant Analysis of Groups I, III, IV, V, VI, and VII

| Variable No. | Identification | I and III | IV and V | VI and VII |
|---|---|---|---|---|
| 13 | r-1 %RO | 1 | | |
| 15 | r-1 $\overline{RO}$ | 4 | | |
| 22 | r-2 $\overline{RO}$ | 2 | | |
| 36 | r-4 $\overline{RO}$ | 3 | | |
| 48 | r-6 %RO | | 3 | 1 |
| 55 | r-7 %RO | | | 2 |
| 57 | r-7 $\overline{RO}$ | | | 3 |
| 76 | r-10 %RO | | 1 | |
| 92 | r-12 $\overline{RO}$ | | | |

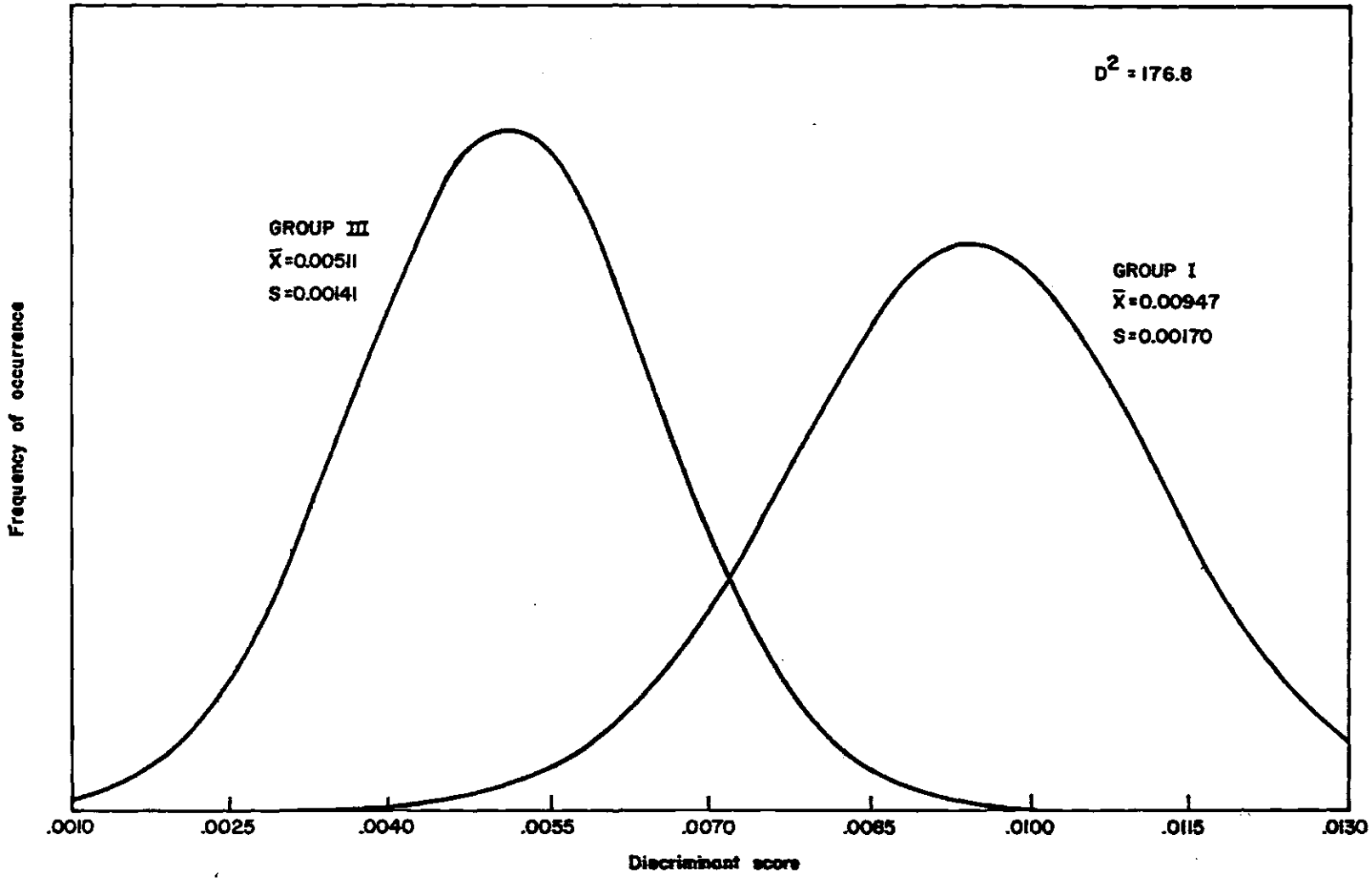
r-i %RO is the percent of precipitation events in range i that produced runoff.
r-i $\overline{RO}$ is the average runoff event from range i

The significant discriminators are again those associated with the smaller rainfall events; however, the association is not as strong as it was in the analysis of length of record. The variables presented

in Table 25 show that as the absolute change in land use decreases from 56 percent in Groups I and III through 28 percent in Groups VI and VII to 16 percent in Groups IV and V, the discriminating variables change from those associated with low rainfall amounts through moderate to higher amounts. There is no apparent reason why this should be the case, and, in fact, there is a good possibility that it was due just to chance because the overall discrimination between Groups IV and V is considerably less than that for Groups I and III. Although all three pairs of values show significant discriminant ability between the groups, the F value shows that Groups I and III are separated in discriminant space much farther than are Groups IV and V or VI and VII.

These conclusions are substantiated by the results of the classification program which show 93 percent correctly-classified observations from Groups I and III, 71 percent for Groups VI and VII, and 66 percent for Groups IV and V. Fifty percent could be expected by chance alone.

The one-dimension discriminant space dispersions for the three pairs of groups are shown in Figs. 17, 18, and 19 for Groups I and III, IV and V, and VI and VII, respectively. The distributions were assumed to be normal. A $\chi^2$ test for conformity was performed on Groups IV and V to check this assumption. Results of this test, confirming the assumption, are presented in Table 26. The histograms of scores for Groups IV and V are also plotted on Fig. 18 along with the normal distributions. The drawings show very plainly that Groups I and III are separated to a much greater extent than are Groups VI and VII or Groups IV and V. It is also quite obvious from the drawings that the group dispersions are not equal for Groups IV and V, but are very nearly




Fibure 17. Distribution of Discriminant Scores for Groups I and III

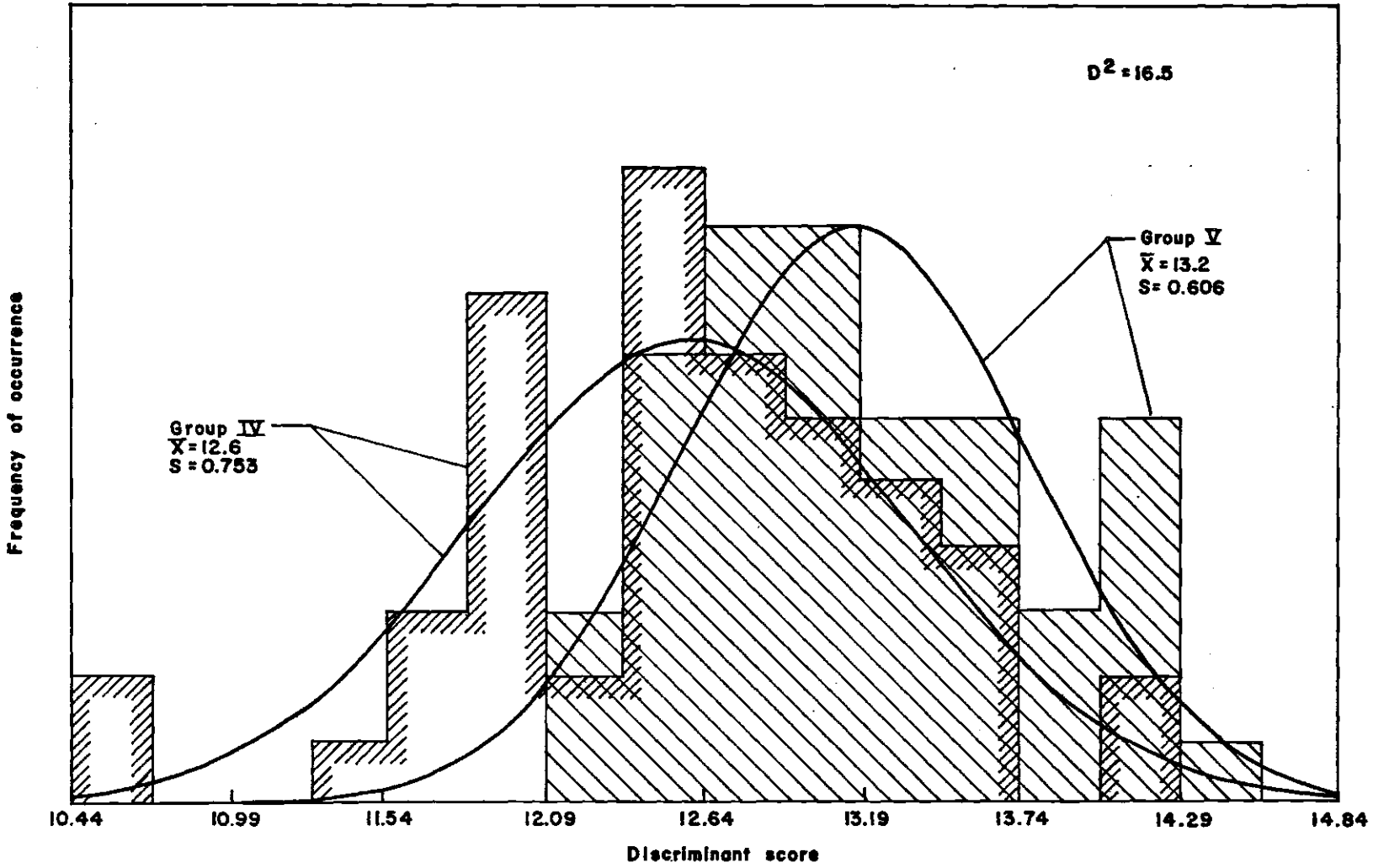



Figure 18. Distribution of Discriminant Scores for Groups IV and V

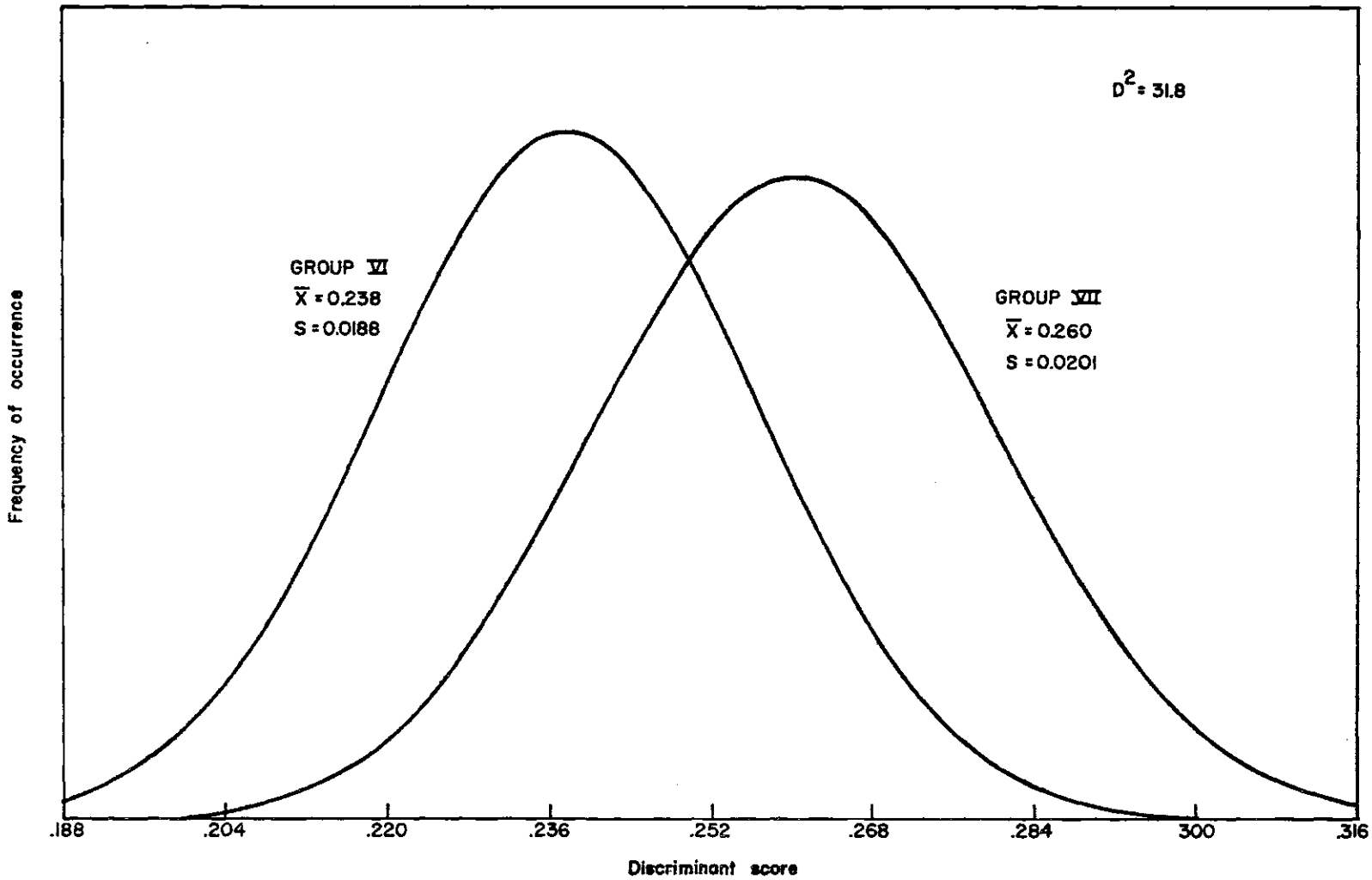



Figure 19. Distribution of Discriminant Scores for Groups VI and VII

identical for Groups I and III, and VI and VII. It is not possible to compare the dispersions of Groups I, IV, and VI which are the groups in which cultivated row crops are prominent, because almost entirely different discriminators were used in each of the three sets.

Table 26. $\chi^2$ Test for Normalcy of Discriminant Scores in Groups IV and V

| Group | Number of Observations | ndf | $\chi^2$ | $\chi^2_{(0.05)}$ |
|---|---|---|---|---|
| IV | 50 | 10 | 5.00 | 18.31 |
| V | 50 | 10 | 10.64 | 18.31 |

It appears from analysis of the synthetic data that the percent of absolute change from one land use to another is more important than is the percent of the watershed in the land uses in which a change takes place. The results would also indicate that moderate to extreme land use changes have a singificant enough impact on the hydrologic characteristics of the watershed model that they should be considered in an analysis in which a change is anticipated, even though a fairly short period of record is used. Again it should be repeated that these results apply only to this watershed model. However, it can be said that these conclusions will probably also be applicable to other areas in the Blacklands where the model is used and in which these crops are grown. It is, however, impossible to estimate the size range of watersheds over which they could apply because this was not investigated in the study.

In general it can be said that these results are logical and somewhat anticipated. Therefore, it is probable that on most watersheds

extreme land use changes will be significant enough to consider, especially when the analyses are based on fairly long periods of record. Small absolute changes in land use will probably not be significant in the analyses of hydrologic data, and in the same light only extreme absolute changes may be significant in the analyses of data from a short period of record.

## The Watershed Model

The watershed model and the technique used to develop synthetic inputs to it were not a primary objective of the research. It was used to generate data for the discriminant analyses such that the effect of short term differences in climatic characteristics could be eliminated. It was also designed to generate data with the same statistical characteristics that the historic period had.

The model used has a good rational basis and was a functional relationship combined with a threshold concept. It is however applicable only for small watersheds in the Blacklands area of Texas where the crops are the same as those described. A technique similar to the one introduced in Chapter IV for incorporating a probabilistic element into a watershed model could and probably should be used on other watershed models. A large number of tests were performed to test the adequacy of the watershed model including its probabilistic element. All of the tests showed that the model would generate synthetic sequences of daily storm runoff statistically similar to the historic record.

It should be pointed out that even though the effect of climatic change can be removed by generating many periods of record under each

land use, the model may still be biased because only short periods of record were available for calibrating the five land use components of the system. This bias is, however, minimal because the four-year period of record, 1954-57, used in the calibration had one very dry year, one very wet year and an average for the period of only about 3 inches below normal. The individual calibrations were tested on a year, 1938, in which the precipitation record was normal.

## Generation of Rainfall and Evaporation

The scheme used to generate inputs to the watershed model, i.e., the occurrence and amount of rainfall and the average monthly evaporation, has been described in detail in Chapter V. The temporal distribution or occurrence of rainfall was represented by a two-state Markov chain of transition probabilities. Tests of the system presented in Tables 10 and 11 show that both the number and distribution of dry days are highly satisfactory. In addition, the average number of dry days per month checked statistically with the historic record as shown in Table 12.

Most rainfall in the area of the country where the watershed is located comes primarily from thunderstorm-type activity and individual storm amounts were found to be uncorrelated with either the previous rainfall or number of days since the previous rainfall. Therefore, the size of the event was assumed to be independent and was generated by random selection from the size distribution of monthly events as defined by the parameters and equations presented in Chapter V. Natural logarithms of amount were generated for two reasons: (1) They were

nearly normally distributed, and (2) negative precipitation values would not be generated. The distribution of the logarithms of precipitation values was a skewed normal distribution with a transform applied to extremely large events. Kolmogorov-Smirnov tests of the distribution by months, presented in Table 17, showed that the scheme was highly satisfactory.

Both the temporal and size generating schemes were then combined and several synthetic 30-year sequences generated. Tests of the results showed that the annual average and all but two of the months were the same as the historical record at the 95 percent conficence level. The other two months were borderline cases. This test was an extremely stringent test because it is very easy to have minor differences in the system which can cause cumulative results to deviate radically from the observed record. Since in this case the long periods were statistically sound, the distribution of individual events can be assumed very good.

As an additional test of the scheme, the distribution of maximum rainfall events in the entire number of years generated was checked against the historical record. In all, 11,430 years of data were generated and all events greater than 5 inches recorded. The maximum event in this period was 34.26 inches. The distribution of these events is plotted on extreme value probability paper in Fig. 20. The points representing the historical record are also shown on the figure. The distribution of events less than 5 inches could not be presented for the synthetic series, however the results presented in Table 17 verify the adequacy of the lower part of the curve.

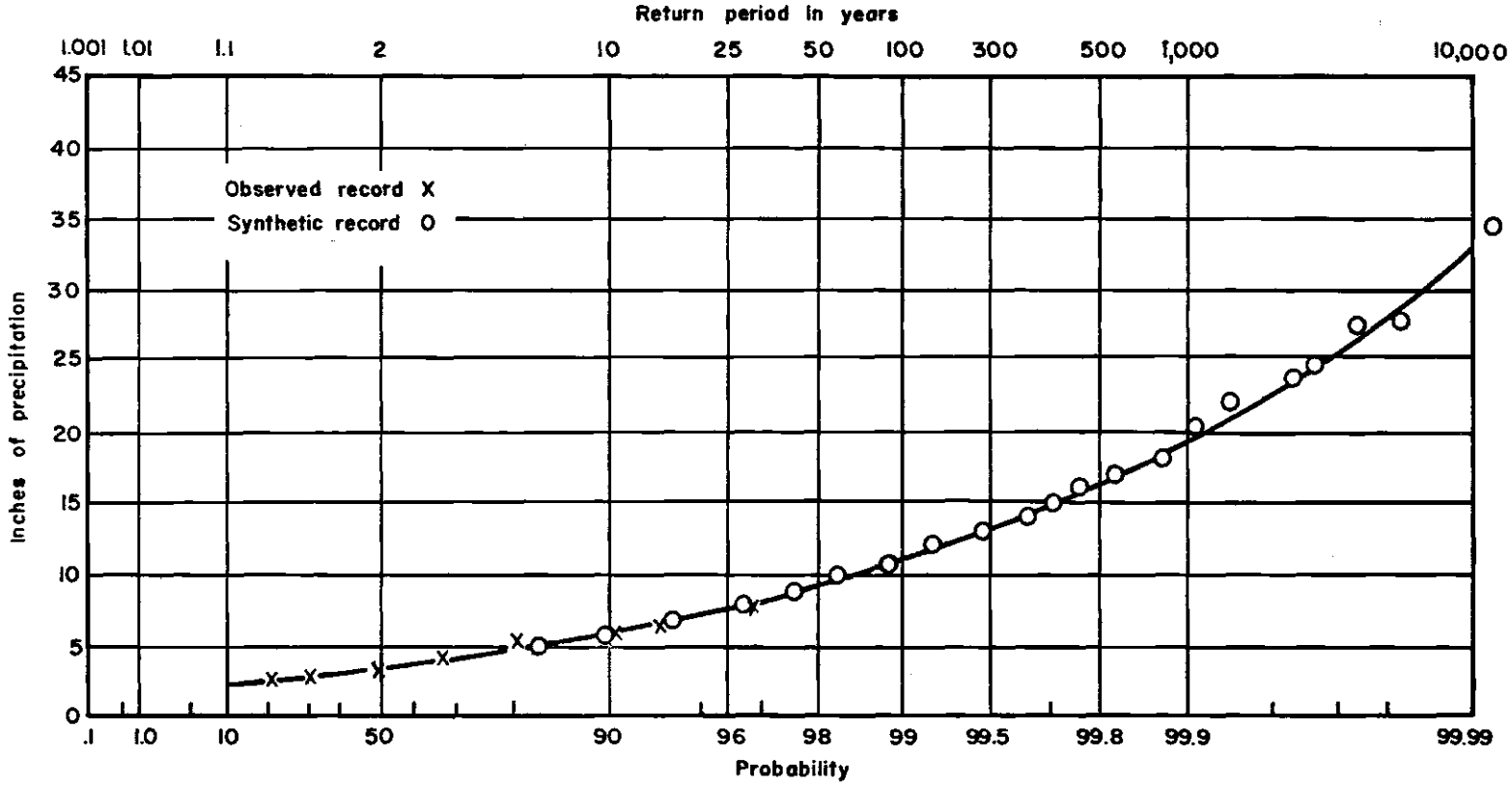



Figure 20. Distribution of Extreme Rainfall Events

At first glance, it appears that the curve may be too high and that the events hould not be as large as those generated. However, the following quotation from Reference 96 would tend to substantiate the results:

> Over most of the area the average annual rainfall is from 35 to 40 inches and at the southwestern limit of the main Blackland prairie is about 30 inches. Short storms of very high intensities are common, particularly during the spring and summer months. Storms of longer duration and large amounts of rainfall occur less frequently. Vance and Lowry list 33 major storms for Texas during the 43 years from 1891 to 1933. In all but 5 of these storms the maximum depth of rainfall was more than 10 inches; in 12 it was more than 15 inches; and in 5, more than 20 inches. Most of these storms covered parts of the Blacklands and some of them centered there.

The curve of Fig. 20 would indicate rainfall amounts of 20 inches or more would occur once in less than 1,200 years and those greater than 11 inches, less often than once in 100 years. This would seem to be in agreement with the above quotation. This check would further indicate that the transform used on extremely large events is operating adequately.

It should be noted that this discussion of extreme rainfall events can be used only to describe the performance of the generation system and not for the purpose of estimating the size of extreme events. This is because the distributions used in the generation scheme were fitted to truncated precipitation distributions of about 30 years duration. Any extrapolation beyond 30 years would be of questionable value. The synthetic 30-year sequences can however be used to study the possible distributions and patterns of rainfall that might be expected.

It is believed that in general the scheme used to generate the

distribution and amount of rainfall is satisfactory and with proper adjustment for amounts, etc. could be applied to a large part of the plains area of the United States where most rainfall occurs as thunderstorm activity producing independent events.

# CHAPTER VIII

## CONCLUSIONS

Conclusions presented in this chapter have been taken from Chapter VII and from Chapter A-III in the Appendix. Because of the physical similarity of the runoff process in most areas, some of the conclusions developed from this study can be considered general; however, all conclusions apply only to small modeled watersheds of about two square miles.

Multiple discriminant analysis can be used to distinguish group differences and find variables which contribute most significantly to group separation.

Components analysis and varimax rotation used in conjunction with a dummy variable can be used to isolate the most significant variables when two or three groups are being considered. In the two-group case, the order of selection can also be determined by this method.

Brier and Allen tests substantiated the use of a change in Mahalanobis' $D^2$ as a test statistic for use in the stepwise selection of predictors and of a $\chi^2$ test on the eigenvalues or roots of $W^{-1}B$ for the number of discriminant functions.

When changes in the land use components of a model are extensive, irrespective of period of record, the significant discriminators are characteristics of runoff from small rainfall events, generally rainfall less than 1 inch.

Extensive changes in the land use components of watershed models will probably be distinguishable in summaries of moderate to long periods of record, but may not be distinguishable on short periods of record. In the same sense, small changes in land use will probably not be significant on short periods of record.

With reference to a watershed model, the effect of a qualitative change in land use cannot be evaluated in general because of the many combinations of soil, climate, and land use; however, it can be stated that in general the percent absolute change in land use is more important than is the percent of the watershed which is in the land use in which a change takes place.

In areas such as the Southern Plains where there is no correlation between consecutive rainfall events, a two-state Markov chain of transition probabilities can be used to describe the temporal distribution of events.

The distribution of size of rainfall events was found to be a skewed log normal distribution with a transform on extremely large events.

# CHAPTER IX

## RECOMMENDATIONS

Results of this study are encouraging enough to show that discriminant analyses can be used in the analysis of hydrologic data. During the course of the research program, a number of problems have become evident along with possible alternative solutions. However, it was not possible to pursue them.

One of the important variables in water resources investigations is the size of the watershed in question. Both the distribution of runoff in time and space as well as the volume of runoff are functions of the drainage area. A possibility for future study would be the development of a watershed model in which both land use and drainage area are variables such that the significance of land use change can be correlated with drainage area.

It might also be of interest to investigate the effect that different watershed models have on the significance of land use change. This would be primarily a function of the degree of specification in the model. A model that completely defined the system would always be able to produce data attributable to the input land use as there would be no probabilistic component to mask the effects of land use change. Other areas of the country, climates, and land uses could also be investigated ultimately leading to a regionalization approach to defining the significance of land use change.

The distribution of size of precipitation events was approximated with a skewed log normal distribution with a transform applied to extremely large events. This was highly satisfactory for the study but could possibly be improved by studying the daily patterns of other rain gages in the vicinity which have longer periods of record. The author was not able to find any theoretical distribution which would perform satisfactorily, however with a longer period of record to work with, it may be possible to find one.

Another somewhat similar problem was the method referred to as the Sammon's approach for incorporating skew into the normal distribution without using tabled values. Since this work was completed, personnel associated with the Hydrologic Center of the Corps of Engineers in Sacramento, California have started using a slightly refined equation attributable to Fiering (102). If investigations in which skewed distributions are to be used were to be duplicated, it would probably be advisable to incorporate the refined equation.

In addition to these recommendations, there are many possible applications for discriminant analysis in the study of hydrology and water resources.

One of the most important uses of discriminant analysis may be in regionalization of hydrologic data. Several variables such as precipitation, intensity of precipitation, runoff, geology, land use, and other characteristics of the runoff such as frequency and duration, all of which are considered to be of importance in hydrologic work could be used as discriminators. The stream gaging stations of the country could then be grouped into proposed regions and then analyzed

by multiple discriminant analysis. The results of classification following the discriminant analysis could be used to reassign the stations. By repeating this process a few times and studying the geographical location of the watersheds, groups which are similar because of their hydrologic characteristics would be developed. It would then be possible to study these watersheds and find why they are similar or even to further subdivide the regions on the basis of additional discriminators such as degree of pollution or salinity. These hydrologically similar groups should be areas in which predictive-type equations developed from one watershed could be extended to other watersheds.

# APPENDIX

## CHAPTER A-1

## RANDOM NUMBER GENERATOR

Data used for the discriminant analyses described in this report were synthetically generated. The analytic models used to generate these data were dependent upon a number generator which would produce long sequences of independent random numbers.

The random number generator available in the IBM Scientific Subroutine Package was tested for length of sequence and was found to repeat after approximately 7,000 numbers. $\chi^2$ tests of the distribution of the random numbers also showed that it was not uniform. These results are in accordance with results of tests by MacLaren and Marsaglia (103) and Van Gelder (104). They tested for uniformity of singles, pairs, and triples; distribution; mean; variance; and autocorrelation of different sample sizes and found that of all the random number generators using standard methods that were tested, none gave satisfactory results.

Van Gelder found several power residue (sometimes called congruential or multiplicative) generators which performed well in his tests. The tests which he used differed somewhat from those of MacLaren and Marsaglia but were found to be compatible with theirs. Of the several methods found to be satisfactory, one, a power residue method with a recycle period of over one-half billion numbers, was selected for use

in the study. Random numbers are generated by the equation

$$U_{n+1} = X\ U_n \pmod{10^d} \qquad \text{(A1-1)}$$

where d is the significant number of digits in the computer, equal to 10 on the one used in this study; X is a constant equal to 100003; and $U_n$ and $U_{n+1}$ are consecutive integers.

Both chi-square tests for distribution and tests for serial correlation were performed on numbers from the generator. Results indicated that the random numbers were independent and uniformly distributed at 95 percent confidence limits.

Random normal numbers were calculated from the uniform numbers by using the direct method, see MacLaren and Marsaglia (103).

$$X_1 = (-2 \ln U_1)^{1/2} \cos 2 \pi U_2 \qquad \text{(A1-2)}$$

where $U_1$ and $U_2$ are random numbers and $X_1$ is a random normal number.

A second method, the "sum of uniform deviates" was not used because it requires at least six times as many uniform numbers as the direct method.

# CHAPTER A-II

# KOLMOGOROV-SMIRNOV TESTS

## The Kolmogorov-Smirnov One-Sample Test

The Kolmogorov-Smirnov one-sample test is a nonparametric goodness of fit test relating to the cumulative distribution function. It is in general more powerful than the chi-square test used for the same purpose because it is based on the maximum difference between the distributions rather than the cumulative difference. The test statistic is

$$D_n = \underset{-\infty < x < \infty}{\text{Max}} \left| F_n(X) - F_o(X) \right| \qquad \text{(A2-1)}$$

in which $F_o(X)$ is the cumulative population distribution being tested, and $F_n(X)$ is the cumulative empirical distribution being tested against. Equation A2-1 is the test statistic for a two-tailed test because $D_n$ is the maximum value of the absolute difference between the distributions.

For very large samples, i.e., $n > 50$, the critical value of $D_n$, $D_\alpha$, is inversely proportional to the square root of the sample size. For $\alpha$ equal to .05; $D_\alpha = 1.36/\sqrt{n}$ where n is the sample size.

The test is useful in hydrologic testing because the distribution of $D_n$ is independent of the form of F(X). The distributions of $D_n$ have been derived and are available in tables; Lindgren and McElrath (105) and Hoel (106).

The test is often used to test discrete population distributions although it is based on the continuous distribution. It has been found that it is on the "safe" side because the actual significance of the resulting test is no bigger than the one assumed in using the tables.

## The Kolmogorov-Smirnov Two-Sample Test

The Kolmogorov-Smirnov one-sample test may be extended to test whether two samples of the same or different sizes are from populations with the same distribution function. The test statistic is

$$D_n = \underset{-\infty < x < \infty}{\text{Max}} \left| G_o(X) - F_o(X) \right| \qquad \text{(A2-2)}$$

in which $G_o(X)$ and $F_o(X)$ are independent distributions assumed to be from the same population. For large sample sizes, the critical value of $D_n$, $D_\alpha$, is inversely proportioned to the square root of n/2 if the samples are both of size n. If the samples are not equal in size, but both are large, $D_\alpha$ is inversely proportional to the square root of $\frac{n_1 n_2}{n_1+n_2}$ where $n_1$ and $n_2$ are the two sample sizes. For $\alpha$ equal to 0.05, the statistic is

$$D_{.05} = \frac{1.36}{\sqrt{\frac{n_1 n_2}{n_1+n_2}}} \qquad \text{(A2-3)}$$

The two-sample test was used in this study in all cases because

the "true" distribution was not known for any of the distributions. The observed distribution is only a sample and, with reference to hydrologic data, can be quite different from the "true" distribution.

# CHAPTER A-III

# MULTIPLE DISCRIMINANT ANALYSES

## Introduction

In the Introduction and in Chapter VI a description of the study was presented. In this chapter the details of the discriminant analysis when applied to the data are presented.

Following is an outline of the material presented in this chapter. Before starting to read the material, it would be advisable to study the outline by first looking at the main points and then the subheadings. You will notice that many of the procedures and tests that are used in the latter parts of the chapter are established, described, and tested in the first part of the chapter where Groups I, II, and III with 30-year summary periods are presented. The tests and procedures are oriented toward 5 objectives: (1) Selecting a reduced set of predictors or variables, the number of which can be feasibly handled in the discriminant analysis computer program, (2) selecting the significant predictors, (3) determining the significant roots or discriminant equations, (4) observing group classification, and (5) determining the significance of the discrimination.

## Part One

I. Thirty-year summary period

   A. Selecting a feasible number of predictors for use in the discriminant analyses computer programs

1. Use of principal components analysis

B. Selecting significant variables

1. Instability from over prediction (too many predictors)

2. Stepwise selection of variables

a. Stability of prediction using the optimum number of predictors

C. Determining the significant roots or discriminant equations

1. Using the $\chi^2$ test

2. Confirmation based on Brier and Allen test

3. Group classification

D. Determining the significance of discrimination

1. Using Wilks' $\Lambda$ test and confirmation with Mahalanobis' $D^2$

2. Testing the equality of group dispersions

E. Summary of discriminant functions for Groups I, II, and III for the 30-year summary period

II. Ten-year summary period

A. Selecting significant variables

B. Determining the significant roots

1. Using the $\chi^2$ test

2. Confirmation with Brier and Allen test

3. Group classification

C. Determining the significance of discrimination

D. Summary of discriminant functions for Groups I, II, and III for the 10-year summary period

III. Two-year summary period

A. Selecting significant variables

1. Using the $\chi^2$ criterion and confirmation with the Brier and Allen test
2. Group classification

B. Determining the significance of the root

C. Determining the significance of discrimination

D. Summary of discriminant functions for Groups I, II, and III for the two-year summary period

Part Two

I. Design of remainder of study

II. Groups I and III

A. Selecting significant variables

1. Primary selection and ranking by component analysis
2. Verification of ranking and selection of significant predictors by the stepwise procedure

B. Group classification

C. Determining the significance of discrimination

D. Summary of discriminant function for Groups I and III

III. Groups IV and V

A. Selecting significant variables

B. Group classification

C. Significance of discrimination

D. Summary of discriminant function for Groups IV and V

IV. Groups VI and VII

A. Selecting significant variables

B. Group classification

C. Significance of discrimination

D. Summary of discriminant function for Groups VI and VII

PART ONE

## Thirty-Year Summary Period

In this part of the report, the effect of length of record on land use group discrimination is considered. The first summary period is 30 years. The land use combinations used correspond to Groups I, II, and III in Table 23, Chapter VI. These are the land use patterns characteristic of the years 1937, 1961, and 1966. The computer facility available for use in this study was an IBM 1130. The computer storage available was adequate to handle only 20 of the 58 variables or discriminants of each observation in the discriminant analysis, therefore the dimensionality of the problem had to be reduced. This is the subject of the next section.

### Selecting a Feasible Number of Variables for Use in the Discriminant Analyses Computer Programs

Use of Principal Components Analysis. Wallis (37) in studying two groups at a time, found that by using a dummy variable, he could very efficiently use principal components analysis and varimax rotation to select the significant variables from the total number of variables used in the discrimination. Publications describing the mathematics and use of components analysis are included in the list of "Other References." The dummy variable used to distinguish group membership is based on group frequencies and was shown by Anderson (5) to give results identical to those of the linear discriminant function. The numerical value of the dummy variable which is orthogonal with respect

to each of the two groups is calculated by

$$X_{d1} = \frac{N_2}{N_1 + N_2} \qquad (A3\text{-}1)$$

$$(d = 1, \cdots, N_1)$$

and

$$X_{d2} = \frac{-N_1}{N_1 + N_2} \qquad (A3\text{-}2)$$

$$(d = 1, \cdots, N_2)$$

in which $N_1$ and $N_2$ are the number of observations in Groups I and II respectively, and $X_{d1}$ and $X_{d2}$ are the dummy variates for Groups I and II respectively. Wallis, using the dummy variable compared five different methods of selecting predictors or variables for the two groups: (1) All variables, (2) stepwise selection, (3) unrelated measurement, (4) reduced rank, and (5) factor score method. The last three are based on components analysis and varimax rotation. He found that the unrelated measurement and reduced rank methods were the most stable. The unrelated measurement method selected all variables which had high loadings on significant factors of the rotated factor weight matrix. The reduced rank method is the same as the unrelated measurement method except that if there is more than one factor-defining variable per factor, then the most significant variable is selected. Fig. 2 in the article by Wallis shows that of the two methods of selecting variables,

the reduced rank method produced the smaller percent of misclassifications when used in forming a discriminant function from small and intermediate sample sizes.

The concept of a single dummy variable as described by Anderson cannot be extended to a three-group case because it is not possible to define orthogonality in three dimensions with a single variate. Therefore, the groups were considered two at a time using the reduced rank method of selecting variables from the rotated factor weight matrix of the two groups. All three possible combinations of two groups, i.e., I and II, I and III, and II and III were analyzed by the same method, and a composite set of variables were selected for use in the discriminant analysis. The variables selected, 20 in number, were those from the factors which had the highest loadings on the dummy variate using only one variable per factor.

Components analysis of a correlation matrix consists of finding all the significant roots and vectors of the matrix by considering all variables at one time. However, the computer available for this work was limited to a maximum of 25 variables; therefore, the data set of 58 variables was divided into three sets for the initial selection of variables. The most significant variables of these three sets were then combined in a fourth set for the final selection by components analysis. Table A1 lists the 58 variables, the sets into which they fall, and the number assigned to them. Components analyses and varimax rotation using this pattern were made for each of the three pairs of groups. Results for the fourth or compositing set for Groups I and II, I and III, and II and III are shown in Tables A2, A3, and A4, respectively.

Table A1. Variables Used in Multiple Discriminant Analysis

| | No. | Variable | No. | Variable | No. | Variable |
|---|---|---|---|---|---|---|
| Set I | 1 | Total No. RO Events | 34 | r-4 % RO | 76 | r-10 % RO |
| | 2 | $\overline{RO}$ | 41 | r-5 " | 83 | r-11 " |
| | 3 | $s_{RO}$ | 48 | r-6 " | 90 | r-12 " |
| | 13 | r-1 % RO | 55 | r-7 " | 97 | r-13 " |
| | 20 | r-2 " | 62 | r-8 " | 104 | r-14 " |
| | 27 | r-3 " | 69 | r-9 " | 111 | r-15 " |
| Set II | 15 | r-1 $\overline{RO}$ | 71 | r-9 " | 119 116* | > .2" $\Delta t$ |
| | 22 | r-2 " | 78 | r-10 " | 120 117* | > .4" " |
| | 29 | r-3 " | 85 | r-11 " | 121 118* | > .7" " |
| | 36 | r-4 " | 92 | r-12 " | 122 119* | >1.0" " |
| | 43 | r-5 " | 99 | r-13 " | 123 120* | >2.0" " |
| | 50 | r-6 " | 106 | r-14 " | 124 121* | >5.0" " |
| | 57 | r-7 " | 113 | r-15 " | | |
| | 64 | r-8 " | 118 115* | >.1" $\Delta t$ | | |
| Set III | 16 | r-1 $s_{RO}$ | 58 | r-7 $s_{RO}$ | 100 | r-13 $s_{RO}$ |
| | 23 | r-2 " | 65 | r-8 " | 107 | r-14 " |
| | 30 | r-3 " | 72 | r-9 " | 114 | r-15 " |
| | 37 | r-4 " | 79 | r-10 " | 115# | AS $\overline{RO}$ |
| | 44 | r-5 " | 86 | r-11 " | 116# | AS $s_{RO}$ |
| | 51 | r-6 " | 93 | r-12 " | 117# | AS g |

*Numbering system for 2-year and 5-year summaries
#Not present in 2-year and 5-year summaries

$\overline{RO}$ is the average runoff for the period
$s_{RO}$ is the standard deviation of runoff for the period
r-i % RO is the percent of precipitation events in range i that produced runoff
r-i $\overline{RO}$ is the average runoff event from range i
r-i $s_{RO}$ is the standard deviation of runoff events in range i
>i" $\Delta t$ is the average length of period between runoff events greater than i inches
AS $\overline{RO}$ is the mean of the annual series of maximum events
AS $s_{RO}$ is the standard deviation of the annual series of maximum events
AS g is the skew coefficient of the annual series of maximum events

Table A2. Varimax Rotated Factor Weight Matrix, Groups I and II for 30-Year Summary Level

| Element No. | Variable | Factor 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | -0.960 | | | | | | | | | | | | | | | | |
| 2 | 2 | 0.716 | | | | | | | | | | | | | | | | | |
| 3 | 3 | | | | 0.909 | | | | | | | | | | | | | | |
| 4 | 13 | | -0.954 | | | | | | | | | | | | | | | | |
| 5 | 15 | | | | | | | | | | | 0.966 | | | | | | | |
| 6 | 20 | | -0.894 | | | | | | | | | | | | | | | | |
| 7 | 23 | | | | | | | | | | 0.958 | | | | | | | | |
| 8 | 65 | | | | | | | | | | | | 0.955 | | | | | | |
| 9 | 69 | | | 0.973 | | | | | | | | | | | | | | | |
| 10 | 76 | | | | | | | | | 0.969 | | | | | | | | | |
| 11 | 83 | | | | | | | -0.976 | | | | | | | | | | | |
| 12 | 90 | | | | | | 0.983 | | | | | | | | | | | | |
| 13 | 92 | | | | | -0.973 | | | | | | | | | | | | | |
| 14 | 106 | | | | | | | | -0.982 | | | | | | | | | | |
| 15 | 113 | | | | 0.958 | | | | | | | | | | | | | | |
| 16 | 115 | 0.548 | | | | | | | | | | | | | -0.728 | | | | |
| 17 | 123 | -0.933 | | | | | | | | | | | | | | | | | |
| 18 | Dummy | 0.167 | 0.205 | 0.225 | 0.131 | -0.145 | 0.121 | -0.117 | -0.132 | 0.116 | -0.076 | -0.096 | 0.121 | -0.866 | -0.047 | -0.003 | -0.001 | 0.005 | 0.000 |

Table A3. Varimax Rotated Factor Weight Matrix, Groups I and III for 30-Year Summary Level

| Element No. | Variable | Factor 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | | | | | | 0.695 | | | | | | | | | | |
| 2 | 13 | | 0.967 | | | | | | | | | | | | | | |
| 3 | 15 | | | | | | | 0.975 | | | | | | | | | |
| 4 | 22 | | | | 0.949 | | | | | | | | | | | | |
| 5 | 23 | | | | 0.960 | | | | | | | | | | | | |
| 6 | 29 | | | | | | | | 0.971 | | | | | | | | |
| 7 | 36 | 0.932 | | | | | | | | | | | | | | | |
| 8 | 37 | 0.930 | | | | | | | | | | | | | | | |
| 9 | 43 | | | 0.938 | | | | | | | | | | | | | |
| 10 | 44 | | | 0.946 | | | | | | | | | | | | | |
| 11 | 71 | | | | | | | | | | | 0.972 | | | | | |
| 12 | 72 | | | | | | | | | | 0.969 | | | | | | |
| 13 | 76 | | | | | -0.987 | | | | | | | | | | | |
| 14 | 115 | | | | | | 0.944 | | | | | | | | | | |
| 15 | 117 | | | | | | | | | | -0.970 | | | | | | |
| 16 | Dummy | -0.166 | 0.760 | -0.131 | -0.222 | 0.079 | -0.063 | 0.164 | -0.108 | 0.131 | -0.157 | 0.071 | 0.482 | -0.006 | 0.001 | 0.019 | 0.001 |

Table A4. Varimax Rotated Factor Weight Matrix, Groups II and III for 30-Year Summary Level

| Element No. | Variable | Factor 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | | | | 0.572 | | | -0.548 | | | | | | | | | | | | |
| 2 | 13 | | | | -0.959 | | | | | | | | | | | | | | | |
| 3 | 15 | | 0.924 | | | | | | | | | | | | | | | | | |
| 4 | 16 | | 0.958 | | | | | | | | | | | | | | | | | |
| 5 | 29 | | | | | | | | | | | | | 0.938 | | | | | | |
| 6 | 36 | | | -0.924 | | | | | | | | | | | | | | | | |
| 7 | 37 | | | -0.930 | | | | | | | | | | | | | | | | |
| 8 | 43 | 0.907 | | | | | | | | | | | | | | | | | | |
| 9 | 44 | 0.934 | | | | | | | | | | | | | | | | | | |
| 10 | 51 | | | | | | | | | | | -0.944 | | | | | | | | |
| 11 | 69 | | | | | | | | | | 0.971 | | | | | | | | | |
| 12 | 76 | | | | | | 0.981 | | | | | | | | | | | | | |
| 13 | 83 | | | | | 0.959 | | | | | | | | | | | | | | |
| 14 | 90 | | | | | | | | | | | | -0.965 | | | | | | | |
| 15 | 92 | | | | | | | | | 0.979 | | | | | | | | | | |
| 16 | 113 | | | | | | | | -0.985 | | | | | | | | | | | |
| 17 | 115 | | | | | | | -0.953 | | | | | | | | | | | | |
| 18 | 117 | | | | | | | | | | | | | | -0.957 | | | | | |
| 19 | Dummy | -0.192 | 0.253 | 0.150 | -0.814 | -0.162 | 0.111 | 0.090 | 0.039 | -0.106 | -0.076 | 0.100 | 0.059 | -0.143 | -0.192 | -0.012 | 0.008 | 0.003 | -0.278 | -0.018 |

All loadings greater than 0.5 are shown. The reduced set, 20 variables, selected from these three tables for use in the discriminant analysis are shown on the composite map of Fig. A1. Variables selected were those which were highly loaded on significant factors. The significant factors are those which were most heavily weighted on the dummy criterion.

Selecting Significant Variables

The first 25 observations, each consisting of the 20-variable set for the three groups, were subjected to the multiple discriminant analysis computer program. Since there were three groups, two discriminant functions were found. Using both discriminant functions and the 20 variables of each observation, the two discriminant scores for each observation were calculated.

Instability from Over Prediction. The two discriminant scores of the first 25 observations of each group, which reduce the test space of 20 variables to the discriminant space of 2 variables, were used with the classification scheme outlined in Chapter III, to check the discriminant ability of the 20 variables. The results of classification based on Rule 2, Equation 42, are shown in Table A5. To check the stability of the discriminant functions, they were used to calculate the discriminant scores of the second 25 observations in each of the three groups. These scores were then used in the classification program. The results of this test are shown in Table A6 again using Rule 2 for group assignment. The fact that a larger percent of the observations in the test set were misclassified than were misclassified in the sample set shows that the discriminant function was overdetermined by including

Figure A1. Composite Map for Groups I, II, III at 30-Year Summary

| Band | $\overline{RO}$ | $s_{RO}$ | %RO | | |
|---|---|---|---|---|---|
| 1 | X | | X | | |
| 2 | X | X | | Total Number of Events | |
| 3 | X | | | Average Runoff | X |
| 4 | X | | | Standard Deviation of Runoff | |
| 5 | X | | | Average of Annual Maximums | X |
| 6 | | | | Standard Deviation of Annual Maximums | |
| 7 | | | | Skew of Annual Maximums | X |
| 8 | | X | | $\Delta t$ for Runoff > 0.1 inch | |
| 9 | | X | X | " " " 0.2 " | |
| 10 | | | X | " " " 0.4 " | |
| 11 | | | X | " " " 0.7 " | |
| 12 | X | | X | " " " 1.0 " | |
| 13 | | | | " " " 2.0 inches | X |
| 14 | X | | | " " " 5.0 " | |
| 15 | X | | | | |

X denotes selection for use in discriminant analysis.

too many variables, some of which were fitting random error of the system.

Table A5. Classification of the Sample Set of 25 Observations from Groups I, II, and III for the 30-Year Summary Level Based on 20 Variables

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 21 | 5 | 0 | 26 |
| | II | 2 | 20 | 0 | 22 |
| | III | 2 | 0 | 25 | 27 |
| | Total | 25 | 25 | 25 | 75 |

Percent Hits - 88 Assignment based on Rule II

Table A6. Classification of the Independent Sample Set of 25 Observations from Groups I, II, and III for the 30-Year Summary Level Based on 20 Variables

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 15 | 5 | 1 | 21 |
| | II | 7 | 20 | 1 | 28 |
| | III | 3 | 0 | 23 | 26 |
| | Total | 25 | 25 | 25 | 75 |

Percent Hits - 77 Assignment based on Rule II

Stepwise Selection of Variables. Using Mahalanobis' $D^2$ as the criterion, the stepwise selection of variables as outlined in Chapter III was used to find the significant variables in the 20 shown on Fig. A1. All 50 observations were used in the selection. Table A7 is a summary of the selection listing the first 8 variables in the order in which they were selected, the test criterion, and the critical $\chi^2$ value. The null hypothesis states that the variable does not contribute a significant amount of information to the discrimination between groups.

The level of significance $\alpha^*$ selected for rejection of the null hypothesis was 0.05.

Table A7. Stepwise Selection of Variables for Groups I, II, and III for the 30-year Summary Level Based on 50 Observations

| Order of Variables Selected | Mahalanobis' $D^2$ | $\Delta D^2$ | Test Criterion $\chi^2(0.05)$ | Accept the Null Hypothesis |
|---|---|---|---|---|
| 13 | 201.17 | 201.17 | 12.00 | * |
| 2 | 258.64 | 57.47 | 11.88 | * |
| 15 | 308.36 | 49.72 | 11.79 | * |
| 76 | 348.32 | 39.96 | 11.66 | * |
| 22 | 375.64 | 27.32 | 11.54 | * |
| 117 | 390.30 | 14.66 | 11.41 | * |
| 69 | 401.10 | 10.80 | 11.27 | |
| 90 | 411.87 | 10.77 | 11.13 | |

The value of the test criterion used in this analysis is a function of the total number of variables. Since the variables are in a sense artificial, almost any number could have been generated, i.e., P is not a fixed number. Therefore, for purposes of the test, the number of possible predictors, P in Equation 30, was set at 20. This corresponds to the maximum number that could be handled by the discriminant analysis program.

Stability of Prediction Using the Optimum Number of Predictors: A stepwise selection of variables based on the first 25 observations of each group gave the same selection as shown on Table A7. These 25 observations were then used in the discriminant analysis program to calculate the discriminant functions.

Using the discriminant functions, the test space of six variables was reduced to the discriminant space of two variables. The

first 25 observations in each of the three groups were then used in the classification program to test the discriminant ability of the 6 variables. The results based on Rule II are as shown in Table A8. The same discriminant equations were used to calculate the scores on the second set of 25 observations, an independent sample, then these were used in the classification program. The results based on Rule II for group classification, are shown in Table A9. The stability of the discriminant functions based on the reduced number of variables is indicated by the fact that classification of the independent sample was just as good (even better) as was the classification of the sample set. These tests verify the stability of the discriminant function based upon a group of variables selected by the stepwise procedure. No additional tests on independent data have been made.

Table A8. Classification of the Sample Set of 25 Observations from Groups I, II, and III for the 30-Year Summary Level Based on 6 Predictors

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 15 | 11 | 1 | 27 |
| | II | 7 | 14 | 0 | 21 |
| | III | 3 | 0 | 24 | 27 |
| | Total | 25 | 25 | 25 | 75 |

Percent of Hits - 71 Assignment based on Rule II

Table A9. Classification of the Independent Sample Set of 25 Observations from Groups I, II, and III for the 30-Year Summary Level Based on 6 Predictors

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 16 | 6 | 0 | 22 |
| | II | 6 | 19 | 0 | 25 |
| | III | 3 | 0 | 25 | 28 |
| | Total | 25 | 25 | 25 | 75 |

Percent of Hits - 80 Assignment based on Rule II

Determining the Significant Roots or Discriminant Equations

Using the $\chi^2$ Test. The 6 variables selected by the stepwise procedure were used with all 50 sets of observations from Groups I, II, and III to develop the two discriminant functions. The latent roots of these two functions and the $\chi^2$ test described in Chapter III are shown in Table A10. According to the test, the largest root, 2.905 was statistically significant at the 95 percent confidence level, but the second root, 0.0625 was not. Since the second root was almost big enough to be significant, an independent test based on the Brier and Allen $\overline{P}$ test statistic described in Chapter III was conducted. It is especially useful because it is based on the difference in predictive ability of two systems.

Table A10. Significance of the Discriminant Function $\chi^2$ Approximations for Groups I, II and III with 50 Observations for the 30-Year Summary Level and 6 Variables

| Function | Root | ndf# | $\chi^2$ | $\chi^2_{(0.05)}$ |
|---|---|---|---|---|
| I | 2.905 | 7 | 198.2 | 16.01 |
| II | 0.0625 | 5 | 8.80 | 12.83 |

# ndf is the number of degrees of freedom (R+G-2j)

Confirmation Based on Brier and Allen Test. Discriminant functions based on one and two roots respectively were developed from the 50 observation sets using the 6 variables. Results of classification based on one and two roots and Rule II for group assignment are presented in Table A11. The $\overline{P}$ score for each of the two systems is also presented. Both the percent of hits and $\overline{P}$ score show that the classification scheme based on two discriminant functions is superior

to that based on one function even though the second root was shown to be insignificant. A test of the significance of the difference in the two $\bar{P}$ values is presented in Table A12. Prediction scheme A is based on one root, scheme B is based on two roots. The calculated t value was 1.70. In order for the two schemes to be statistically significant at the 95 percent confidence level, the t value should exceed 1.96. On this basis, the null hypothesis which states that the two prediction schemes are equal, could not be rejected. Thus classification based on one discriminant function as indicated by the test on the size of the roots was upheld by the Brier and Allen test.

Table A11. Classification of Groups I, II, and III for the 30-Year Summary Level Using One and Two Discriminant Functions Respectively

One Function

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 21 | 11 | 1 | 33 |
| | II | 23 | 39 | 0 | 62 |
| | III | 6 | 0 | 49 | 55 |
| | Total | 50 | 50 | 50 | 150 |

Percent Hits = 73; $\bar{P}$ = 0.343; Assignment based on Rule II

Two Functions

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 27 | 9 | 1 | 37 |
| | II | 17 | 41 | 0 | 58 |
| | III | 6 | 0 | 49 | 55 |
| | Total | 50 | 50 | 50 | 150 |

Percent Hits = 78; $\bar{P}$ = 0.317; Assignment based on Rule II

Table A12. Test of the Significance of Including the Second Root in a Classification Scheme Based on Brier and Allen Scores for the 30-Year Summary Level

| $\Sigma P_A$ | $\Sigma P_B$ | $\Sigma(P_{Ai}-P_{Bi})^2$ | ndf | t | $t_{(0.05)}$ |
|---|---|---|---|---|---|
| 51.455 | 47.597 | 5.2479 | 149 | 1.70 | 1.96 |

A - one root; B - two roots

Group Classification. Group classification based on the optimum number of variables and functions is the first one shown in Table A11.

Determining the Significance of Discrimination

Using the Wilks' Λ Test and Confirmation with Mahalanobis' $D^2$. The overall discriminating power of the 6 variables can be tested using either the Wilks' Λ test described in Chapter III or the Mahalanobis' $D^2$ test as defined by Equations 27 and 28. The values of Λ and $D^2$ based on the 6 variables are .241 and 388.76, respectively. The value of F calculated from Λ by the procedure described in Chapter III is 24.54. The critical value of F at the 95 percent confidence level is 1.79. Therefore, the null hypothesis, $H_2$, which states that the variables do not discriminate between the groups, i.e., that the centroids of the three groups are the same, is rejected. The critical value of Mahalanobis' $D^2$ is 21.03 based on 12 degrees of freedom and a 95 percent confidence level. Therefore, the null hypothesis is also rejected by this test. Since the two tests are emperically equal, the F test results will be presented in the balance of the report.

Testing the Equality of Group Dispersions. The feasibility of the test of $H_2$ above, is based on the hypothesis, $H_1$, that the dispersion matrices are equal. The test statistic described in Chapter III

for making this test was calculated from the same data set used to test the equality of dispersions. The value of M, the test statistic, was 86.77 giving an F value of 1.95 with a critical value of F at the 95 percent confidence level of 1.38. The null hypothesis that the dispersions of the three groups are equal is therefore rejected. The results of testing both $H_1$ and $H_2$ are shown in Table A13.

Table A13. Testing the Hypotheses $H_1$ and $H_2$ for Groups I, II, and III for the 30-Year Summary Level and 6 Variables

| | Test Statistic | ndf# Numerator $f_1$ | Denominator $f_2$ | $F_{f_2}^{f_1}$ | $F_{f_2}^{f_1}(0.05)$ |
|---|---|---|---|---|---|
| $H_2$ | $\Lambda$ = .241 | 12 | 286 | 24.71 | 1.79 |
| $H_1$ | M = 86.77 | 42 | 64,152 | 1.95 | 1.38 |

# ndf is the number of degrees of freedom

Since the hypothesis $H_1$ was rejected, the significance of the tests of hypothesis $H_2$ may be questionable. However, since the test for $H_2$ was so significant, and the fact that moderate departures from homogenity of dispersions do not affect the test of $H_2$, it is likely that there is a statistically significant degree of discrimination between the three groups of data.

Summary of Discriminant Functions for Groups I, II, and III for the 30-Year Summary Period

Using the results of the stepwise selection of variables and the tests on the number of roots, optimum discrimination for Groups I, II, and III is based on six variables and one discriminant function. The discriminant function which is a linear combination of the six variables

is presented in vector format as $V_{jp}$ in Table A14. The tabled values correspond to the coefficients, $V_{jp}$, in Equation 2 (j=1; p=1, ···, 6). The variable number corresponds to the variable numbers in Table A1. To show the relative discriminative contribution of the variables in the discriminant function, the normalized vector, $V_{jp}$, is adjusted by multiplying corresponding elements by the square roots of the diagonal elements of the pooled within-groups matrix W. The scaled vector, $V_{jp}$*, is also presented in Table A14. It can be seen from the size of the scaled coefficients, $V_{jp}$*, that the variables contribute to the discriminant function in an order of significance equivalent to the order in which they were selected.

Also shown on Table A14 are the group means and standard deviations in the test space, the group centroids and dispersions in discriminant space, and the W and B matrices. The total correlation matrix, also shown, shows that the six variables selected are relatively independent.

## Ten-Year Summary Period

One of the objectives of the study was to determine the effect that period of record has on the ability to distinguish differences in hydrologic characteristics due to a land use change. The data for evaluating this objective were obtained for Groups I, II, and III at the same time that the data used in the analyses described thus far were generated. As mentioned in Chapter VI, 50 synthetic records summarized at 2-, 5-, 10-, 20-, and 30-year intervals were collected. Each of these observations consisted of the 58 variables listed in

Table A14. Characteristics of Groups I, II, and III in the Test and Discriminant Spaces at Optimum Solution for the 30-Year Summary Period

| | Test Space Summaries | | | | | | Discriminant Function | |
|---|---|---|---|---|---|---|---|---|
| | Group I | | Group II | | Group III | | | |
| Variable Number | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. | Normalized $v_{jp}$ | Scaled $v_{jp}^*$ |
| 13 | 19.30 | 2.89 | 20.76 | 2.13 | 14.24 | 2.11 | -.000458 | -.0134 |
| 2 | .239 | .0275 | .218 | .0301 | .275 | .0356 | .0219 | .0083 |
| 15 | .00427 | .00057 | .00447 | .00062 | .00395 | .00051 | -.983 | -.0068 |
| 76 | 97.69 | 5.79 | 95.13 | 7.43 | 98.85 | 3.30 | .0000784 | .0055 |
| 22 | .0102 | .00246 | .0107 | .00213 | .0115 | .00208 | .182 | .0049 |
| 117 | .0828 | .568 | .142 | .474 | -.239 | .404 | -.000690 | -.0041 |

Discriminant Space Summaries

| Group I | | Group II | | Group III | |
|---|---|---|---|---|---|
| Centroid | Dispersion | Centroid | Dispersion | Centroid | Dispersion |
| .00164 | .0000026 | .000160 | .0000016 | .00560 | .0000012 |

Pooled W Matrix

| Variable Number | 13 | 2 | 15 | 76 | 22 | 117 |
|---|---|---|---|---|---|---|
| 13 | 849.997 | -.928172 | .002919 | 499.664 | .223853 | -2.48060 |
| 2 | -.928172 | .143526 | .000868 | -1.62205 | .001817 | .247753 |
| 15 | .002919 | .000868 | .000048 | -.010232 | .000029 | .000688 |
| 76 | 499.664 | -1.62205 | -.010232 | 4879.82 | .131487 | -17.6895 |
| 22 | .223853 | .001817 | .000029 | .131487 | .000732 | -.001040 |
| 117 | -2.48060 | .247753 | .000688 | -17.6895 | -.001040 | 34.8064 |

B Matrix

| Variable Number | 13 | 2 | 15 | 76 | 22 | 117 |
|---|---|---|---|---|---|---|
| 13 | 1171.19 | -9.62808 | .088793 | -563.951 | -.191408 | 70.0314 |
| 2 | -9.62808 | .081029 | -.000751 | 5.04858 | .001425 | -.569479 |
| 15 | .088793 | -.000751 | .000007 | -.047375 | -.000013 | .005239 |
| 76 | -563.951 | 5.04858 | -.047375 | 362.127 | .059677 | -32.3561 |
| 22 | -.191408 | .001425 | -.000013 | .059677 | .000043 | -.011939 |
| 117 | 70.0314 | -.569479 | .005239 | -32.3561 | -.011939 | 4.20834 |

Total Correlation Matrix

| Variable Number | 13 | 2 | 15 | 76 | 22 | 117 |
|---|---|---|---|---|---|---|
| 13 | 1.000 | -.496 | .275 | -.020 | .026 | .241 |
| 2 | -.496 | 1.000 | .033 | .100 | .246 | -.109 |
| 15 | .275 | .033 | 1.000 | -.107 | .080 | .128 |
| 76 | -.020 | .100 | -.107 | 1.000 | .095 | -.111 |
| 22 | .026 | .246 | .080 | .095 | 1.000 | -.075 |
| 117 | .241 | -.109 | .128 | -.111 | -.075 | 1.000 |

Table A1. The 2- and 5-year summary records were not long enough to use in determining the characteristics of the annual maximum series, therefore the three variables characterizing the series were not included in the records. Since each of these records were summaries of a progressively longer period of time, they were used to show what happens to the ability to discriminate as the length of record changes.

The analysis which follows describes the results at the 10- and 2-year summary periods respectively. It was found that there was a statistically significant degree of discrimination at the 2-year level. Although the ability to discriminate at this level was considerably reduced from that at the 10- and 30-year levels, the differences were not considered significant enough to warrant the time involved in analyzing the intermediate levels at 5 and 20 years.

Selecting Significant Variables

The data set for the 10-year summary level of Groups I, II, and III consisted of 50 observations on each of the 58 variables in Table A1. The components analysis scheme using a dummy variable, described in the analysis of the 30-year summary data, was used with this data to reduce the number of variables to a maximum of 20. Results of components analysis and varimax rotation for the three pairs of land use groups are presented in Tables A15, A16, and A17. As in the previous presentation, all factor loadings greater than 0.5 are shown. The 20 variables selected from these tables for use in the stepwise selection of variables are shown on the composite map of Fig. A2.

The stepwise approach to the selection of variables, described previously, was used to find the significant variables in the 20 shown

Table A15. Varimax Rotated Factor Weight Matrix, Groups I and II for the 10-Year Summary Level

| Element No. | Variable | Factor 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | | | -0.921 | | | | | | | | | | |
| 2 | 55 | | | | | | | 0.984 | | | | | | |
| 3 | 29 | | -0.927 | | | | | | | | | | | |
| 4 | 50 | 0.939 | | | | | | | | | | | | |
| 5 | 85 | | | | | | -0.983 | | | | | | | |
| 6 | 92 | | | | | | | | -0.968 | | | | | |
| 7 | 99 | | | | | | | | | | 0.981 | | | |
| 8 | 118 | | | | 0.967 | | | | | | | | | |
| 9 | 30 | | -0.957 | | | | | | | | | | | |
| 10 | 51 | 0.951 | | | | | | | | | | | | |
| 11 | 72 | | | | | | | | | 0.974 | | | | |
| 12 | 86 | | | | | 0.978 | | | | | | | | |
| 13 | Dummy | 0.167 | -0.074 | -0.044 | 0.081 | 0.087 | 0.074 | 0.080 | -0.123 | 0.066 | 0.110 | -0.952 | 0.002 | 0.001 |

Table A16. Varimax Rotated Factor Weight Matrix, Groups I and III for the 10-Year Summary Level

| Element No. | Variable | Factor 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | | | | | | | | | | | | | | -0.862 | | | |
| 2 | 13 | | | | | | | | | | | | | | | 0.703 | | |
| 3 | 27 | -0.916 | | | | | | | | | | | | | | | | |
| 4 | 34 | | | | | | | | | | | | -0.931 | | | | | |
| 5 | 15 | | | -0.950 | | | | | | | | | | | | | | |
| 6 | 22 | | | | | | | | | 0.982 | | | | | | | | |
| 7 | 36 | | | | | | | | -0.971 | | | | | | | | | |
| 8 | 57 | | | | | | 0.988 | | | | | | | | | | | |
| 9 | 99 | | | | 0.982 | | | | | | | | | | | | | |
| 10 | 118 | | | | | | | | | | | | | 0.842 | | | | |
| 11 | 124 | | -0.950 | | | | | | | | | | | | | | | |
| 12 | 16 | | | -0.950 | | | | | | | | | | | | | | |
| 13 | 30 | | | | | | | | | | -0.982 | | | | | | | |
| 14 | 65 | | | | | -0.978 | | | | | | | | | | | | |
| 15 | 72 | | | | | | | 0.986 | | | | | | | | | | |
| 16 | 117 | | | | | | | | | | | -0.960 | | | | | | |
| 17 | Dummy | -0.405 | -0.082 | -0.149 | 0.130 | 0.114 | -0.091 | 0.115 | 0.108 | -0.094 | -0.088 | -0.197 | -0.192 | -0.014 | 0.163 | 0.163 | -0.787 | 0.000 |

Table A17. Varimax Rotated Factor Weight Matrix, Groups II and III for the 10-Year Summary Level

| Element No. | Variable | Factor 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | | | | | | | | | | | | | | | | | | | | | 0.662 | | | |
| 2 | 13 | | | | | | | | | | | | | | | | | | | | | | 0.732 | | |
| 3 | 27 | -0.910 | | | | | | | | | | | | | | | | -0.901 | | | | | | | |
| 4 | 34 | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | 76 | | | | | | 0.979 | | | | | | | | | | | | | | | | | | |
| 6 | 90 | | | | | | | | | 0.973 | | | | | | | | | | | | | | | |
| 7 | 15 | | | | | | | | | | | | -0.935 | | | | | | | | | | | | |
| 8 | 22 | | | | | -0.978 | | | | | | | | | | | | | | | | | | | |
| 9 | 29 | | | | | | | | | | | | | | 0.964 | | | | | | | | | | |
| 10 | 36 | | | 0.934 | | | | | | | | | | | | | | | | | | | | | |
| 11 | 50 | | 0.881 | | | | | | | | | | | | | | | | | | | | | | |
| 12 | 57 | | | | | | | 0.962 | | | | | | | | | | | | | | | | | |
| 13 | 92 | | | | | | | | 0.981 | | | | | | | | | | | | | | | | |
| 14 | 113 | | | | | | | | | | | | | | | | | | | -0.860 | | | | | |
| 15 | 118 | | | | | | | | | | | | | | | | | | -0.829 | | | | | | |
| 16 | 124 | | | | | | | | | | | | | 0.974 | | | | | | | | | | | |
| 17 | 26 | -0.940 | | | | | | | | | | | | | | | | | | | | | | | |
| 18 | 37 | | | 0.947 | | | | | | | | | | | | | | | | | | | | | |
| 19 | 44 | | | | | | | | | | | 0.983 | | | | | | | | | | | | | |
| 20 | 51 | | 0.921 | | | | | | | | | | | | | | | | | | | | | | |
| 21 | 65 | | | | | | | | | | | | | | | -0.957 | | | | | | | | | |
| 22 | 116 | | | | -0.954 | | | | | | | | | | | | | | | | | | | | |
| 23 | 117 | | | | | | | | | | -0.950 | | | | | | | | | | | | | | |
| 24 | Dummy | -0.457 | -0.179 | -0.161 | 0.082 | 0.064 | -0.079 | -0.085 | -0.098 | -0.093 | -0.149 | -0.100 | -0.119 | 0.163 | -0.106 | 0.102 | -0.472 | -0.218 | 0.073 | 0.063 | 0.009 | -0.124 | 0.269 | -0.007 | -0.006 |

Figure A2. Composite Map for Groups I, II, III at 10-Year Summary

| Band | $\overline{RO}$ | $s_{RO}$ | %RO | | |
|---|---|---|---|---|---|
| 1 | X | X | X | Total Number of Events | |
| 2 | | | | Average Runoff | X |
| 3 | X | | X | Standard Deviation of Runoff | |
| 4 | X | | X | Average of Annual Maximums | |
| 5 | | X | | Standard Deviation of Annual Maximums | |
| 6 | X | X | | Skew of Annual Maximums | X |
| 7 | | | X | Δt for Runoff > 0.1 inch | X |
| 8 | | X | | " " " 0.2 " | |
| 9 | | X | | " " " 0.4 " | |
| 10 | | | | " " " 0.7 " | |
| 11 | | X | | " " " 1.0 " | |
| 12 | X | | | " " " 2.0 inches | |
| 13 | X | | | " " " 5.0 " | X |
| 14 | | | | | |
| 15 | | | | | |

X denotes selection for use in discriminant analysis

on Fig. A2. A summary of the selection listing the variables in the order in which they were selected, the test criterion, and the critical $\chi^2$ value are presented in Table A18. It is interesting to note that when variable 99 was added to the set of variables, the increase in $D^2$ was 5.46. When variable 117 was added to this same set (not shown), the increase in $D^2$ was only 4.23. Yet when variable 117 was added after the inclusion of variable 99, the increase was 8.50. This emphasizes a statement by Tatsuoka and Tiedeman (12) that in certain cases, variables which in themselves do not significantly differentiate between two groups may nevertheless enhance the discriminatory power of other variables. These "covariance" variables show up in some of the other stepwise selection analyses.

Table A18. Stepwise Selection of Variables for Groups I, II, and III for the 10-Year Summary Level

| Order of Variables Selected | Mahalanobi's $D^2$ | $\Delta D^2$ | Test Criterion $\chi^2(0.05)$ | Accept the Null Hypothesis |
|---|---|---|---|---|
| 27 | 87.74 | 87.74 | 12.00 | * |
| 2 | 113.82 | 26.08 | 11.88 | * |
| 13 | 128.41 | 14.59 | 11.79 | * |
| 65 | 143.41 | 15.00 | 11.66 | * |
| 99 | 148.87 | 5.46 | 11.54 | |
| 117 | 157.37 | 8.50 | 11.41 | |
| 16 | 160.95 | 3.58 | 11.27 | |

Determining the Significant Roots

Using the $\chi^2$ Test. The four variables selected by the stepwise procedure were used to develop the two discriminant functions. The latent roots of these two functions and the $\chi^2$ test described previously are shown in Table A19. The tests show that the largest root,

1.020, was statistically significant at the 95 percent confidence level but that the smaller root, 0.0263, was not. As with the 30-year summary data, the Brier and Allen $\overline{P}$ test was conducted to check the significance of the root.

Table A19. Significance of the Discriminant Function $X^2$ Approximations for Groups I, II, and III with 50 Observations for the 10-Year Summary Level and 4 Variables

| Function | Root | ndf | $X^2$ | $X^2_{(0.05)}$ |
|---|---|---|---|---|
| I | 1.020 | 5 | 103.00 | 12.83 |
| II | 0.0263 | 3 | 3.81 | 9.35 |

Confirmation with Brier and Allen Test. Group assignment probabilities needed to carry out the Brier and Allen test were calculated from the discriminant functions based on one and two roots. Results of classification based on one and two roots and Rule II for group assignment are presented in Table A20. The $\overline{P}$ score for each of the two systems is also presented on Table A20. The classification results show, as would be expected, better results for the two-root solution as opposed to the one-root solution. The test on the significance of the difference of the two $\overline{P}$ values is presented in Table A21. The calculated t value was .96 and the critical t value at the 95 percent confidence level is 1.96. Therefore, the null hypothesis could not be rejected, and discrimination based on two roots is statistically no better than that based on one root. Again this is in agreement with the $X^2$ test on root size.

Table A20. Classification of Groups I, II, and III for the 10-Year Summary Level Using One and Two Discriminant Functions, Respectively

One Function

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 14 | 17 | 7 | 38 |
| | II | 29 | 28 | 0 | 57 |
| | III | 7 | 5 | 43 | 55 |
| | Total | 50 | 50 | 50 | 150 |

Percent Hits - 57; $\overline{P}$ = .471; Assignment based on Rule II

Two Functions

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 28 | 16 | 6 | 50 |
| | II | 16 | 28 | 2 | 46 |
| | III | 6 | 6 | 42 | 54 |
| | Total | 50 | 50 | 50 | 150 |

Percent Hits - 65; $\overline{P}$ = .461; Assignment based on Rule II

Table A21. Test of the Significance of Including the Second Root in a Calculation Scheme Based on Brier and Allen Scores for the 10-Year Summary Level

| $\Sigma P_A$ | $\Sigma P_B$ | $\Sigma(P_{Ai}-P_{Bi})^2$ | ndf | t | $t_{(0.05)}$ |
|---|---|---|---|---|---|
| 70.689 | 69.132 | 2.6451 | 149 | .96 | 1.96 |

A - one root; B - two roots

Group Classification. Group classification based on the optimum number of variables and functions is the first one shown in Table A20 for one function.

Determining the Significance of Discrimination. Tests of the hypotheses $H_2$ and $H_1$ on the overall discriminating power of the four variables, and of the equality of group dispersions as described in

detail for the 30-year summary are presented in Table A22. The test on $H_2$ shows a highly significant degree of discrimination between the three groups. The equality of dispersions, $H_1$, is not rejected thus the dispersions can be assumed equal. Since $H_1$ is not rejected, the test of $H_2$ can be assumed to be statistically sound.

Table A22. Testing the Hypotheses $H_2$ and $H_1$ for Groups I, II, and III for the 10-Year Summary Level and 4 Variables

| | Test Statistic | ndf Numerator $f_1$ | ndf Denominator $f_2$ | $F^{f_1}_{f_2}$ | $F^{f_1}_{f_2}(0.05)$ |
|---|---|---|---|---|---|
| $H_2$ | $\Lambda$ = .4822 | 8 | 289 | 15.90 | 1.96 |
| $H_1$ | M = 25.77 | 20 | 77,566 | 1.24 | 1.57 |

Summary of Discriminant Functions for Groups I, II, and III at the 10-Year Level

Using results of the stepwise selection of variables and one root, optimum discrimination for Groups I, II, and III is based on four variables and one discriminant function. The discriminant function is presented in vector format, $V_{jp}$, in Table A23. Variables correspond to the numbers in Table A1. The scaled vector which shows the relative contribution of the four variables is presented in Table A23 as $V_{jp}$*. Also shown on Table A23 are the group means and standard deviations in the test space, the group centroids and dispersions in discriminant space, the W and B matrices, and the total correlation matrix.

## Two-Year Summary Period

### Selecting Significant Variables

The data set for the 2-year summary level of Groups I, II, and

Table A23. Characteristics of Groups I, II, and III in the Test and Discriminant Spaces at Optimum Solution for the 10-Year Summary Level

| | Test Space Summaries | | | | | | Discriminant Function | |
|---|---|---|---|---|---|---|---|---|
| | Group I | | Group II | | Group III | | | |
| Variable Number | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. | Normalized $V_{jp}$ | Scaled $V_{jp}$* |
| 27 | 58.86 | 10.29 | 57.77 | 10.28 | 42.14 | 9.34 | .00974 | 1.179 |
| 2 | .236 | .0581 | .217 | .0467 | .278 | .0594 | -.972 | -.649 |
| 13 | 20.03 | 4.11 | 20.61 | 3.91 | 14.40 | 3.14 | .0193 | .878 |
| 65 | .459 | .250 | .419 | .257 | .548 | .259 | -.235 | -.729 |

Discriminant Space Summaries

| Group I | | Group II | | Group III | |
|---|---|---|---|---|---|
| Centroid | Dispersion | Centroid | Dispersion | Centroid | Dispersion |
| .623 | .0308 | .652 | .0302 | .289 | .0201 |

Pooled W Matrix

| Variable Number | 27 | 2 | 13 | 65 |
|---|---|---|---|---|
| 27 | 14,653.0 | -2.55061 | 2,880.06 | 75.8952 |
| 2 | -2.55061 | .445662 | -5.17116 | .178335 |
| 13 | 2,880.06 | -5.17116 | 2,060.32 | 22.0391 |
| 65 | 75.8952 | .178335 | 22.0391 | 9.60196 |

B Matrix

| Variable Number | 27 | 2 | 13 | 65 |
|---|---|---|---|---|
| 27 | 8,748.31 | -27.3142 | 3,177.67 | -57.6346 |
| 2 | -27.3142 | .097805 | -10.4717 | .206801 |
| 13 | 3,177.67 | -10.4717 | 1,178.42 | -22.1144 |
| 65 | -57.6346 | .206801 | -22.1144 | .437273 |

Total Correlation Matrix

| Variable Number | 27 | 2 | 13 | 65 |
|---|---|---|---|---|
| 27 | 1.000 | -.265 | .696 | .038 |
| 2 | -.265 | 1.000 | -.373 | .165 |
| 13 | .696 | -.373 | 1.000 | -.000 |
| 65 | .038 | .165 | -.000 | 1.000 |

Figure A3. Composite Map for Groups I, II, III at 2-Year Summary

| Band | $\overline{RO}$ | $s_{RO}$ | %RO | |
|---|---|---|---|---|
| 1 | | | X | Total Number of Events |
| 2 | X | | X | Average Runoff |
| 3 | X | | X | Standard Deviation of Runoff |
| 4 | | X | X | Average of Annual Maximums |
| 5 | X | | | Standard Deviation of Annual Maximums |
| 6 | | X | | Skew of Annual Maximums |
| 7 | X | X | | Δt for Runoff > 0.1 inch |
| 8 | X | X | | " " " 0.2 " |
| 9 | | | | " " " 0.4 " X |
| 10 | | | X | " " " 0.7 " |
| 11 | | | X | " " " 1.0 " |
| 12 | X | | | " " " 2.0 inches |
| 13 | X | | | " " " 5.0 " |
| 14 | | X | | |
| 15 | | | | |

X denotes selection for use in discriminant analysis

III consisted of 50 observations on each of the 55 variables in Table A1. Results of the components analysis and varimax rotation of the 2-year summary data using the dummy variable to reduce the number of variables to 20 are presented for the three pairs of groups in Tables A24, A,25, and A26. As in previous presentations, all factor loadings greater than 0.5 are shown. The 20 variables selected from these three tables for use in the stepwise selection of variables are displayed on the composite map of Fig. A3.

Using the $\chi^2$ Criterion and Confirmation with the Brier and Allen Test. Results of the stepwise selection of variables are presented in Table A27. Listed are the variables in the order of selection, the test criterion, and the critical $\chi^2$ values. Using Mahalanobis' $D^2$ as the test criterion, only one variable would be selected. However, the second variable is almost significant therefore the Brier and Allen $\overline{P}$ test was used to check the significance of the second variable. With one variable, the test and discriminant space are the same, i.e., there is only one root, but with two variables there are two roots. Therefore, in order to test the two-variable case, the significance of both roots was checked. The results are presented in Table A28. The test shows that the first root, 0.188 is significant whereas the second root is not. In the previous two checks on the significance of roots, the Brier and Allen test has confirmed the $D^2$ test, therefore the significance of the second root was not checked and the two-variable solution based on one root was compared to the one-variable one-root solution.

Table A24. Varimax Rotated Factor Weight Matrix, Groups I and II for 2-Year Summary Level

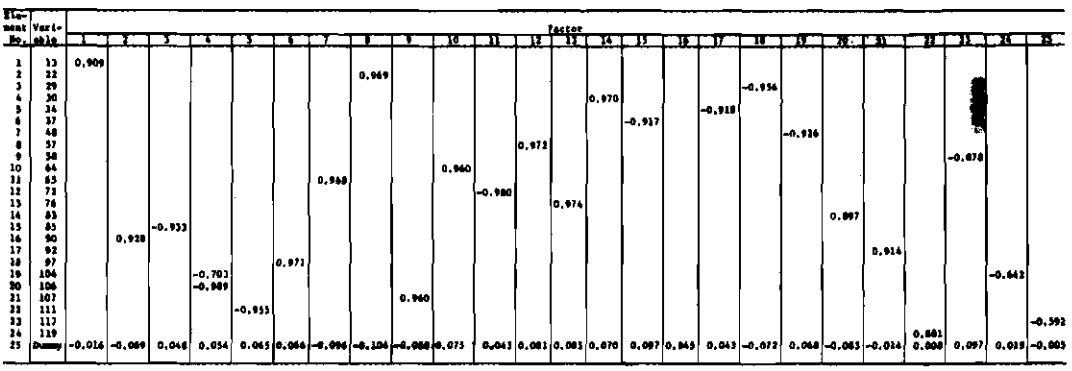


Table A25. Varimax Rotated Factor Weight Matrix, Groups I and III for 2-Year Summary Level

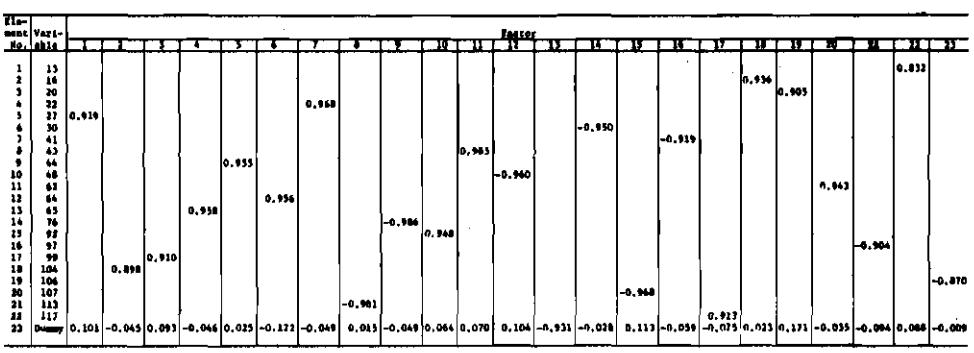


Table A26. Varimax Rotated Factor Weight Matrix, Groups II and III for 2-Year Summary Level

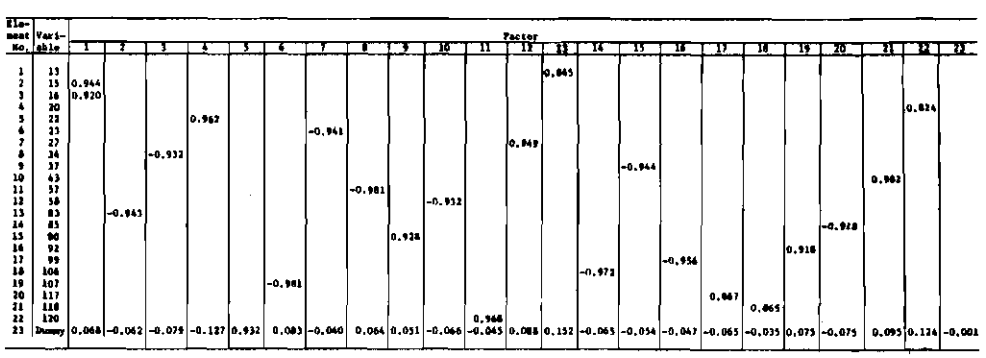


Table A27. Stepwise Selection of Variables for Groups I, II, and III for the 2-Year Summary Level

| Order of Variable Selection | Mahalanobis' $D^2$ | $\Delta D^2$ | Test Criterion $\chi^2(0.05)$ | Accept the Null Hypothesis |
|---|---|---|---|---|
| 20 | 19.84 | 19.84 | 12.00 | * |
| 22 | 30.83 | 10.99 | 11.88 | |
| 29 | 39.19 | 8.36 | 11.79 | |
| 107 | 46.03 | 6.84 | 11.66 | |
| 13 | 53.20 | 7.17 | 11.54 | |
| 64 | 59.22 | 6.02 | 11.41 | |
| 48 | 65.34 | 6.12 | 11.27 | |
| 37 | 69.88 | 4.54 | 11.13 | |

Table A28. Significance of the Discriminant Functions $\chi^2$ Approximations for Groups I, II, and III for the 2-Year Summary Level and 2 Variables

| Function | Root | ndf | $\chi^2$ | $\chi^2(0.05)$ |
|---|---|---|---|---|
| I | .1876 | 3 | 25.35 | 9.35 |
| II | 0.0280 | 1 | 4.08 | 5.02 |

Group assignment probabilities needed to carry out the Brier and Allen test were calculated from the discriminant functions of these two systems. Results of classification by both systems are presented in Table A29. They show that even though the percent hits in the two-variable case is poorer than the one-variable case, the predictive ability on the basis of probabilities is better as indicated by the lower $\overline{P}$ score. The difference between the two $\overline{P}$ values is not significant however, as shown by the data in Table A30. The Brier and Allen test therefore substantiated the $\chi^2$ test for the number of variables to be included in the discriminant function.

Table A29. Classification of Groups I, II, and III for the 2-Year Summary Level Using One and Two Variables, Respectively

One Variable

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 33 | 30 | 14 | 77 |
| | II | 1 | 3 | 0 | 4 |
| | III | 16 | 17 | 36 | 69 |
| | Total | 50 | 50 | 50 | 150 |

Percent Hits - 48; $\overline{P}$ = .614; Assignment based on Rule II

Two Variables

| | | Observed Group | | | |
|---|---|---|---|---|---|
| | | I | II | III | Total |
| Predicted Group | I | 8 | 12 | 7 | 27 |
| | II | 23 | 25 | 6 | 54 |
| | III | 19 | 13 | 37 | 69 |
| | Total | 50 | 50 | 50 | 150 |

Percent Hits - 47; $\overline{P}$ = .604; Assignment based on Rule II

Table A30. Test of the Significance of Including the Second Variable in a Classification Scheme Based on Brier and Allen Scores for the 2-Year Summary Level

| $\Sigma P_A$ | $\Sigma P_B$ | $\Sigma(P_{Ai}-P_{Bi})^2$ | ndf | t | $t_{(0.05)}$ |
|---|---|---|---|---|---|
| 92.174 | 90.562 | 1.6626 | 149 | 1.25 | 1.96 |

A - one variable; B - two variables

Determining the Significance of the Root

The significance of the single root based on one variable was tested as described previously and the results are shown on Table A31. The root is statistically significant.

Table A31. Significance of the Discriminant Functions $\chi^2$ Approximation for Groups I, II, and III for the 2-Year Summary Level and 1 Variable

| Function | Root | ndf | $\chi^2$ | $\chi^2_{(0.05)}$ |
|---|---|---|---|---|
| I | .1359 | 2 | 18.82 | 5.99 |

Determining the Significance of Discrimination

Tests of the hypotheses $H_2$ and $H_1$ on the overall discriminating power of the single variable and of the equality of group dispersions are presented in Table A32. In this case where there is only one variable, the test on the significance of the root is identical to the test of $H_2$, the significance of overall discrimination, and the results are the same, i.e., rejection of the hypothesis that the group centroids are equal. The test of $H_1$ shows that the null hypothesis is not rejected, which means that there is no significant difference between the group dispersions. For this reason, the test of $H_2$ is statistically sound.

Table A32. Testing the Hypotheses $H_2$ and $H_1$ for Groups I, II, and III for the 2-Year Summary Level and 1 Variable

| | Test Statistic | ndf Numerator $f_1$ | ndf Denominator $f_2$ | $F_{f_2}^{f_1}$ | $F_{f_2}^{f_1}(0.05)$ |
|---|---|---|---|---|---|
| $H_2$ | $\Lambda$ = .8804 | 4 | 292 | 9.80 | 2.43 |
| $H_1$ | M = 3.06 | 2 | 48,620 | 1.52 | 3.00 |

Summary of Discriminant Functions for Groups I, II, and III for the 2-Year Summary Period.

The stepwise selection of variables shows that optimum

discrimination for Groups I, II, and III is based on one variable and therefore, on only one discriminant function. The coefficient for the function is presented in Table A33. The variable number corresponds to the variable number in Table A1. Also shown in Table A33 are the group centroids and dispersions in discriminant space. Since only one variable was used in the analysis, the test space and discriminant space are identical and the coefficient in the discriminant function becomes 1.0. The W and B matrices and the correlation matrix are also shown in Table A33.

## PART TWO

### Design of Remainder of Study

Previous discussions show that there is significant discrimination between the three groups, but it is not clear what the differences are. A plot of both discriminant scores for the 50 observations of the three groups at the 30-year summary level was made in order to see graphically the group separation. The plot is shown in Fig. A4. Also shown in the figure are the 50 and 10 percent centours, i.e., centile contours beyond which 50 and 10 percent of the points are expected to lie. These centours were calculated by Equations 37 and 38 in which the X's are the two discriminant space scores. The major axis of each of the ellipses is defined by the bivariate regression line between the two discriminant scores

The plot shows that the three groups are not equally spaced from one another. Group III, Bermuda pasture predominant, is separated from the other two groups; Groups I, cultivated row crops predominant, and

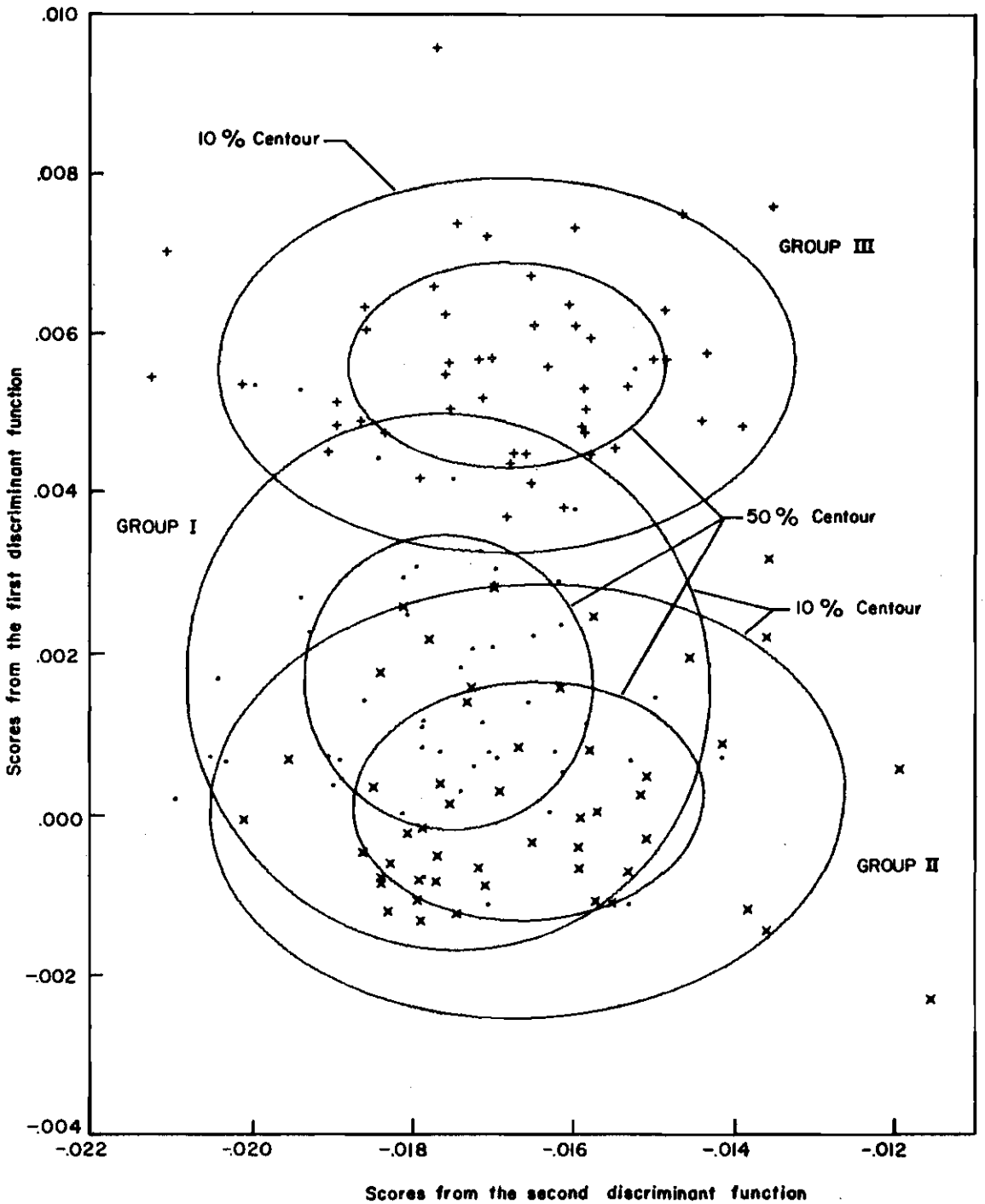



Figure A4. Distribution of the Two Discriminant Scores for Groups I, II, and III for the 30-Year Summary Level

Table A33. Characteristics of Groups I, II, and III in the Test and Discriminant Spaces at Optimum Solution for the 2-Year Summary Level

| Variable Number | Test Space Summaries | | | | | | Discriminant Function | |
|---|---|---|---|---|---|---|---|---|
| | Group I | | Group II | | Group III | | Normalized | Scaled |
| | Mean | Std.Dev. | Mean | Std.Dev. | Mean | Std.Dev. | $V_{jp}$ | $V_{jp}^*$ |
| 20 | 44.62 | 20.73 | 43.71 | 21.79 | 28.72 | 17.11 | 1.000 | 242.2 |

Discriminant Space Summaries

| Group I | | Group II | | Group III | |
|---|---|---|---|---|---|
| Centroid | Dispersion | Centroid | Dispersion | Centroid | Dispersion |
| 44.62 | 429.8 | 43.71 | 475.0 | 28.72 | 292.7 |

Pooled W Matrix

| Variable Number | 20 |
|---|---|
| 20 | 58,674.8 |

B Matrix

| Variable Number | 20 |
|---|---|
| 20 | 7,973.06 |

Total Correlation Matrix

| Variable Number | 20 |
|---|---|
| 20 | 1.000 |

II, native grass meadow predominant, are similar. The difference between Groups I and II would probably have been greater if the percent of the watershed in native grass meadow had been higher in Group II. It was only 49 percent, whereas cultivated row crops were 72 percent when predominant in Group I.

In setting up additional groups to determine the significance of land use change, two things were considered:

(1) Which is more important on a modeled watershed, a nearly complete change in land use on a small part of the watershed, or a partial change on a larger part of the watershed?

(2) How much of a change in land use is necessary in order to show a significant difference in modeled hydrologic characteristics?

Both of these questions must be evaluated in terms of the land use and watershed being modeled. The first question is subjective and can only be answered in relative terms because changing from cultivated row crops to Bermuda pasture, for example, will make more of a change in the hydrologic characteristics than will changing to native grass meadow.

Land use Groups IV through VII were set up in an attempt to answer both of the questions as completely as possible and yet restrict the number of additional groups to four. The restriction on the number of additional runs was made because of the computer time required; approximately 12 hours per set of 50 observations on a land use complex. Since only four additional groups were considered, the groups were set up in pairs with the changes in land use taking place between only two crops. The two land uses selected for study were Bermuda pasture and

cultivated row crops. They were selected for two reasons:

(1) The change from one to the other represents a "middle of the road" change. It is not as extreme as going from Bermuda pasture to native grass meadow nor is it as small as going from cultivated row crops to native grass meadow.

(2) In groups I and III the land use complexes were almost reversed from one another with respect to these two crops. The two crops accounted for approximately 84 percent of the watershed with one or the other representing about 70 percent.

In Groups IV and V, a partial change from one land use to another took place in crops representing a large part of the watershed. The land use complex for these runs are shown on Table 23 in Chapter VI. Bermuda pasture and cultivated row crops accounted for 84 percent of the watershed. In Group IV, Bermuda pasture was on 34 percent of the area and cultivated row crops were on 50 percent of the area. In Group V, these percents were reversed. In Groups VI and VII, a nearly complete change in land use took place on a small part of the watershed. The land use complexes for these groups are also shown on Table 23. Both crops accounted for only 36 percent of the watershed with Bermuda pasture on 4 percent of the watershed and cultivated row crops on 32 percent of the watershed for Group VI. In Group VII these percents were reversed.

In light of the second consideration used in setting up the pattern for Groups IV through VII, the same four groups may be viewed in a different light. In Groups I and III, the amount of change from one land use to the other was about 56 percent. In Groups VI and VII

the change is 28 percent or about one-half of the change in Groups I and III, and in Groups IV and V, the change was 16 percent, a little over one-half the change in Groups VI and VII.

## Groups I and III

### Selecting Significant Variables

Primary Selection and Ranking by Component Analysis. Selection of variables to be used in the discriminant analysis of Groups I and III was based, as in the previous discussions, on the results of components analysis and a varimax rotation of the factor weight matrix. Since, in the two-group case, the dummy variable, when given values according to Equations A3-1 and A3-2, can be used to differentiate between groups, it was surmised that the ranking of variables to be included in the discriminant analysis should correspond to a ranking based on varimax rotation. Results of the components analysis and varimax rotation of the factor weight matrix for Groups I and III were presented in Table A3. Using the data presented in the table, the variables selected for consideration in the multiple discriminant analysis, in decreasing order of importance, are presented in Table A34. Selection was based upon a consideration of all factors with a weight on the dummy criterion of at least 0.10. Since the reduced rank method of selecting variables was being used, only one variable was selected from each of the three factors which had common variables. The three dual loaded factors represented rainfall ranges 2, 4, and 5. The significant variables in each of the three ranges were the means and standard deviations of runoff. Both were nearly equally loaded,

therefore the average runoff was selected because it was felt that it would be more meaningful and was also more highly correlated with the dummy variable.

Table A34. Ordering of Variables for Groups I and III Based on Components Analysis and Varimax Rotation of Factor Weight Matrix

| Variable No. | Factor Weight on Dummy Criterion |
|---|---|
| 13 | .760 |
| 22 | -.222 |
| 36 | -.166 |
| 15 | .164 |
| 117 | -.157 |
| 43 | -.131 |
| 72 | .131 |
| 29 | .108 |

Verification of Ranking and Selection of Significant Predictors by the Stepwise Procedure. To check the ranking of the variables, a stepwise selection was performed. The number of variables considered in the selection were those shown in Table A34. Results of the stepwise selection are presented in Table A35. A comparison of the two lists shows that within the range of significant variables, the order of selection is the same.

On the basis of these results, selection of variables for Groups IV and V, and VI and VII will be based on a combination of factor analysis and stepwise selection. The factor analysis will be used to select the order of variables and the stepwise selection used to find the significant number.

Table A35. Stepwise Selection of Variables for Groups I and III

| Order of Variables Selected | Mahalanobis' $D^2$ | $\Delta D^2$ | Test Criterion $\chi^2(0.05)$ | Accept the Null Hypothesis |
|---|---|---|---|---|
| 13 | 98.98 | 98.98 | 7.48 | * |
| 22 | 148.16 | 49.12 | 7.24 | * |
| 36 | 164.00 | 15.84 | 6.97 | * |
| 15 | 176.86 | 12.86 | 6.63 | * |
| 29 | 181.64 | 4.78 | 6.24 | |
| 43 | 183.86 | 2.22 | 5.73 | |
| 72 | 184.56 | .70 | 5.02 | |
| 117 | 183.86 | -.70 | 3.84 | |

Since this one and the next two analyses are on only two groups at a time, there will be only one discriminant function. Testing the significance of this function is equivalent to testing the hypothesis of difference in group centroids, therefore only the test of $H_2$ and $H_1$ are presented.

Group Classification

The 50 observations on the variables selected for analyzing the Group I and III data were used in the multiple discriminant analysis program to calculate the discriminant function. The function was in turn used to calculate the discriminant scores for probabilistic classification into one or the other of the two groups. Results of classification are presented in Table A36.

Table A36. Classification of Groups I and III

| | | Observed Group I | Observed Group II | Total |
|---|---|---|---|---|
| Predicted Group | I | 45 | 2 | 47 |
| | II | 5 | 48 | 53 |
| | Total | 50 | 50 | 100 |

Percent Hits - 93 Assignment based on Rule II

Determining the Significance of Discrimination

The tests of the hypotheses $H_2$ and $H_1$ on the overall discriminating power of the four variables, and of the equality of group dispersions are presented in Table A37. The test of $H_2$ shows that there is a significant amount of discrimination between the two groups, i.e., the hypothesis of equality of group centroids is rejected. The test of $H_1$, the equality of group dispersions, is not rejected. Thus, the test of $H_2$ is not questionable.

Table A37. Testing the Hypotheses $H_2$ and $H_1$ for Groups I and III

| | Test Statistic | ndf Numerator $f_1$ | ndf Denominator $f_2$ | $F_{f_2}^{f_1}$ | $F_{f_2}^{f_1}(0.05)$ |
|---|---|---|---|---|---|
| $H_2$ | $\Lambda$ = .3397 | 4 | 95 | 46.16 | 2.46 |
| $H_1$ | M = 12.32 | 10 | 45,915 | 1.18 | 1.83 |

Summary of Discriminant Functions for Groups I and III

Coefficients of the discriminant function are presented in Table A38. The variable numbers correspond to those in Table A1. Also shown in Table A38 are the scaled vectors which show the relative contribution of the variables to the discrimination. The group means and standard deviations in the test space and the group centroids and dispersions in discriminant space are presented. Also given on Table A38 are the W and B matrices and the total correlation matrix.

## Groups IV and V

Selecting Significant Variables

Results using components analysis and varimax rotation of the

Table A38. Characteristics of Groups I and III in the Test and Discriminant Spaces at Optimum Solution

| Variable Number | Test Space Summaries | | | | Discriminant Function | |
|---|---|---|---|---|---|---|
| | Group I | | Group III | | Normalized | Scaled |
| | Mean | Std. Dev. | Mean | Std. Dev. | $V_{jp}$ | $V_{jp*}$ |
| 13 | 19.30 | 2.89 | 14.24 | 2.11 | .000635 | .0159 |
| 22 | .0102 | .00246 | .0115 | .00208 | -.384 | -.0087 |
| 36 | .0193 | .00408 | .0218 | .00482 | -.143 | -.0063 |
| 15 | .00427 | .00057 | .00395 | .00051 | .912 | .0049 |

Discriminant Space Summaries

| Group I | | Group III | |
|---|---|---|---|
| Centroid | Dispersion | Centroid | Dispersion |
| .00947 | .0000029 | .00511 | .0000020 |

Pooled W Matrix

| Variable No. | 13 | 22 | 36 | 15 |
|---|---|---|---|---|
| 13 | 627.360 | .218965 | .264135 | .007989 |
| 22 | .218965 | .000509 | .000212 | .000016 |
| 36 | .264135 | .000212 | .001952 | .000054 |
| 15 | .007989 | .000016 | .000054 | .000029 |

B Matrix

| Variable No. | 13 | 22 | 36 | 15 |
|---|---|---|---|---|
| 13 | 640.184 | -.162944 | -.323102 | .040230 |
| 22 | -.162944 | .000041 | .000082 | -.000010 |
| 36 | -.323102 | .000082 | .000163 | -.000020 |
| 15 | .040230 | -.000010 | -.000020 | .000002 |

Total Correlation Matrix

| Variable No. | 13 | 22 | 36 | 15 |
|---|---|---|---|---|
| 13 | 1.000 | .067 | -.036 | .241 |
| 22 | .067 | 1.000 | .273 | .040 |
| 36 | -.036 | .273 | 1.000 | .130 |
| 15 | .241 | .040 | .130 | 1.000 |

Table A39. Varimax Rotated Factor Weight Matrix, Groups IV and V

| Element No. | Variable | Factor 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.955 | | | | | | | | | | | | |
| 2 | 48 | | | | | | 0.979 | | | | | | | |
| 3 | 69 | | | | | | | | | 0.978 | | | | |
| 4 | 76 | | | | 0.983 | | | | | | | | | |
| 5 | 83 | | | | | | | | 0.981 | | | | | |
| 6 | 22 | | | | | | | -0.964 | | | | | | |
| 7 | 43 | | | | | 0.972 | | | | | | | | |
| 8 | 64 | | -0.940 | | | | | | | | | | | |
| 9 | 92 | | | | | | | | | | | -0.932 | | |
| 10 | 118 | -0.894 | | | | | | | | | | | | |
| 11 | 65 | | -0.907 | | | | | | | | | | | |
| 12 | 93 | | | -0.951 | | | | | | | | | | |
| 13 | Dummy | 0.076 | -0.066 | -0.088 | -0.120 | -0.066 | -0.104 | 0.059 | -0.099 | -0.068 | 0.961 | -0.111 | 0.003 | 0.000 |

data from Groups IV and V to select a reduced number of variables are presented in Table A39. As in previous presentations, all factor loadings greater than 0.5 are shown. The variables selected from the table are presented in their order of importance in Table A40. The same criteria as described for selection of variables from Groups I and III were used in this selection. Vector number 10 with a very high loading on the dummy criterion indicates that there is probably very little difference between the groups.

Table A40. Ordering of Variables for Groups IV and V Based on Components Analysis and Varimax Rotation of the Factor Weight Matrix

| Variable No. | Factor Weight on Dummy Criterion |
|---|---|
| 76 | -.120 |
| 92 | -.111 |
| 48 | -.104 |

The stepwise selection of variables from those of Table A40 is shown in Table A41. The test criterion of all three variables is large enough that all three variables are significant. It was found that in considering the next most significant variables, they did not contribute significantly more information to the discrimination between the two groups to be included.

Group Classification

The 50 observations on Variables 76, 92, and 48 were used in the multiple discriminant analysis program to calculate the discriminant function for Groups IV and V. The function was in turn used to calculate the discriminant scores for probabilistic classification into one

or the other of the two groups. Results of the classification are presented in Table A42.

Table A41. Significance of Variables Considered for Discriminant Analysis of Groups IV and V

| Order of Adding Variables | Mahalanobis' $D^2$ | $\Delta D^2$ | Test Criterion $\chi^2_{(0.05)}$ | Accept the Null Hypothesis |
|---|---|---|---|---|
| 76 | 5.82 | 5.82 | 5.73 | * |
| 92 | 12.40 | 6.58 | 5.02 | * |
| 48 | 16.52 | 4.12 | 3.84 | * |

Table A42. Classification of Groups IV and V

| | | Observed Group | | |
|---|---|---|---|---|
| | | IV | V | Total |
| Predicted Group | IV | 28 | 12 | 40 |
| | V | 22 | 38 | 60 |
| | Total | 50 | 50 | 100 |

Percent Hits - 66 Assignment based on Rule II

Significance of Discrimination

Tests of the hypotheses $H_2$ and $H_1$ on the overall discriminating power of the three varialbes and the equality of the group dispersions are presented in Table A43. The test of $H_2$ shows that there is a significant amount of discrimination between the two groups, i.e., the hypothesis of equality of group centroids is rejected. However, the test results are not nearly as strong as the previous tests. The test of $H_1$, the equality of group dispersions is rejected. Therefore, the validity of the test of $H_2$ can be questioned.

Table A43. Testing the Hypotheses $H_2$ and $H_1$ for Groups IV and V

| | Test Statistic | ndf Numerator $f_1$ | ndf Denominator $f_2$ | $F^{f_1}_{f_2}$ | $F^{f_1}_{f_2}(0.05)$ |
|---|---|---|---|---|---|
| $H_2$ | $\Lambda$ = .849 | 3 | 96 | 5.68 | 2.70 |
| $H_1$ | M = 32.59 | 6 | 69,582 | 5.25 | 2.09 |

Summary of the Discriminant Function for Groups IV and V

Coefficients of the discriminant function defined by the three variables are presented in Table A44. The variable numbers correspond to those in Table A1. The scaled vectors which show the relative contribution of the variables to the discrimination are also presented. The group means and standard deviations in the test space and the group centroids and dispersions in discriminant space are shown also. Included in the table are the W and B matrices and the total correlation matrix.

## Groups VI and VII

Selecting Significant Variables

Results of using components analysis and varimax rotation of the data from Groups VI and VII to select a reduced number of variables are presented in Table A45. Factor loadings greater than 0.5 are shown. Variables selected from the table are presented in their order of importance in Table A46. The criterion used in selecting variables from Groups I and III were also used in this selection. Factor 14 is uniquely loaded on the dummy variable. Again this is an indication that the degree of discrimination may be small although not as small as for Groups IV and V.

Table A44. Characteristics of Groups IV and V in the Test and Discriminant Spaces at Optimum Solution

| Variable Number | Test Space Summaries | | | | Discriminant Function | |
|---|---|---|---|---|---|---|
| | Group IV | | Group V | | Normalized | Scaled |
| | Mean | Std. Dev. | Mean | Std. Dev. | $V_{jp}$ | $V_{jp*}$ |
| 76 | 97.09 | 5.44 | 99.15 | 2.58 | .0945 | 3.98 |
| 92 | .767 | .471 | .533 | .477 | -.994 | -4.66 |
| 48 | 77.71 | 6.19 | 80.37 | 6.97 | .0539 | 3.52 |

Discriminant Space Summaries

| Group IV | | Group V | |
|---|---|---|---|
| Centroid | Dispersion | Centroid | Dispersion |
| 12.59 | .567 | 13.17 | .368 |

Pooled W Matrix

| Variable No. | 76 | 92 | 48 |
|---|---|---|---|
| 76 | 1779.51 | 13.1788 | 202.775 |
| 92 | 13.1788 | 22.0206 | 35.6277 |
| 48 | 202.775 | 35.6277 | 4260.80 |

B Matrix

| Variable No. | 76 | 92 | 48 |
|---|---|---|---|
| 76 | 106.859 | -12.0499 | 137.242 |
| 92 | -12.0499 | 1.36064 | -15.4939 |
| 48 | 137.242 | -15.4939 | 176.377 |

Total Correlation Matrix

| Variable No. | 76 | 92 | 48 |
|---|---|---|---|
| 76 | 1.000 | .005 | .118 |
| 92 | .005 | 1.000 | .063 |
| 48 | .118 | .063 | 1.000 |

Table A45. Varimax Rotated Factor Weight Matrix, Groups VI and VII

| Element No. | Variable | Factor 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 34 | | | | | | 0.938 | | | | | | | | | | | | |
| 2 | 41 | | | | | | | | | 0.960 | | | | | | | | | |
| 3 | 48 | | | | | | | | | | | 0.964 | | | | | | | |
| 4 | 55 | | 0.955 | | | | | | | | | | | | | | | | |
| 5 | 62 | | | | | | | | | | 0.981 | | | | | | | | |
| 6 | 69 | | | | | | | | | | | | 0.966 | | | | | | |
| 7 | 76 | | | | | | | | | | | | | -0.978 | | | | | |
| 8 | 29 | 0.949 | | | | | | | | | | | | | | | | | |
| 9 | 57 | | | 0.950 | | | | | | | | | | | | | | | |
| 10 | 85 | | | | | | | | -0.980 | | | | | | | | | | |
| 11 | 113 | | | | -0.939 | | | | | | | | | | | | | | |
| 12 | 118 | | | | | | | | | | | | | | | -0.842 | | | |
| 13 | 16 | | | | | 0.930 | | | | | | | | | | | | | |
| 14 | 30 | 0.967 | | | | | | | | | | | | | | | | | |
| 15 | 58 | | | 0.949 | | | | | | | | | | | | | | | |
| 16 | 100 | | | | | | | 0.977 | | | | | | | | | | | |
| 17 | 117 | | | | | | | | | | | | | | | | 0.862 | | |
| 18 | Dummy | -0.085 | -0.166 | -0.154 | 0.078 | -0.047 | -0.105 | -0.076 | -0.113 | -0.101 | -0.117 | -0.182 | -0.128 | 0.123 | 0.898 | -0.013 | -0.087 | 0.002 | -0.001 |

Table A46. Ordering of Variables for Groups VI and VII Based on Components Analysis and Varimax Rotation of the Factor Weight Matrix

| Variable No. | Factor Weight on Dummy Criterion |
|---|---|
| 48 | -.182 |
| 55 | -.166 |
| 57 | -.154 |
| 69 | -.128 |
| 76 | .122 |
| 62 | -.117 |
| 85 | -.113 |
| 34 | -.105 |
| 41 | -.101 |

Using the variables in Table A46, a stepwise selection of variables was performed. The test criterion shows that three of the variables are statistically significant. The results are presented in Table A47.

Table A47. Significance of Variables Considered for Discriminant Analysis of Groups VI and VII

| Order of Adding Variables | Mahalanobis' $D^2$ | $\Delta D^2$ | Test Criterion $\chi^2(0.05)$ | Accept the Null Hypothesis |
|---|---|---|---|---|
| 48 | 14.33 | 14.33 | 7.70 | * |
| 55 | 23.86 | 11.53 | 7.48 | * |
| 57 | 31.84 | 7.98 | 7.24 | * |
| 69 | 37.97 | 6.13 | 6.97 | |
| 76 | 40.74 | 2.77 | 6.63 | |
| 62 | 45.59 | 5.25 | 6.24 | |

Group Classification

The 50 observations on Variables 48, 55, and 57 were used in the multiple discriminant analysis program to calculate the discriminant function for Groups VI and VII. The function was in turn used to

calculate the discriminant scores for the probabilistic classification into one or the other of the two groups. Results of the classification are presented in Table A48.

Table A48. Classification of Groups VI and VII

| | | Observed Group | | |
|---|---|---|---|---|
| | | VI | VII | Total |
| Predicted Group | VI | 38 | 17 | 55 |
| | VII | 12 | 33 | 45 |
| | Total | 50 | 50 | 100 |

Percent Hits - 71 Assignment based on Rule II

Significance of Discrimination

Tests of the hypotheses $H_2$ and $H_1$ on the overall discriminating power of the three variables and the equality of the group dispersions are presented in Table A49. The test of $H_2$ shows that there is a significant amount of discrimination between the two groups, i.e., the hypothesis of equality of group centroids is rejected. The test of $H_1$, the equality of group dispersions is not rejected. Therefore the test of $H_2$ can be assumed to be statistically sound.

Table A49. Testing the Hypotheses $H_2$ and $H_1$ for Groups VI and VII

| | Test Statistic | ndf Numerator $f_1$ | ndf Denominator $f_2$ | $F_{f_2}^{f_1}$ | $F_{f_2}^{f_1}(0.05)$ |
|---|---|---|---|---|---|
| $H_2$ | $\Lambda$ = .745 | 3 | 96 | 10.93 | 2.70 |
| $H_1$ | M = 2.79 | 6 | 69,582 | .37 | 2.09 |

Summary of the Discriminant Function for Groups VI and VII

Coefficients of the discriminant function in normal and scaled form are presented in Table A50. The scaled vectors show the relative

contribution of the variables to the discrimination. The group means and standard deviations in the test space and the centroids and dispersions in discriminant space are also shown. Presented in Table A50 are the W and B matrices and the total correlation matrix.

Table A50. Characteristics of Groups VI and VII in the Test and Discriminant Spaces at Optimum Solution

| Variable Number | Test Space: Group VI Means | Test Space: Group VI Std. Dev. | Test Space: Group VII Means | Test Space: Group VII Std. Dev. | Discriminant Function: Normalized $V_{jp}$ | Discriminant Function: Scaled $V_{jp^*}$ |
|---|---|---|---|---|---|---|
| 48 | 69.62 | 7.40 | 75.12 | 7.04 | .00167 | .120 |
| 55 | 74.47 | 8.91 | 80.90 | 8.23 | .00125 | .106 |
| 57 | .0285 | .00947 | .0338 | .0107 | 1.000 | .100 |

Discriminant Space Summaries

| Group VI Centroid | Group VI Dispersion | Group VII Centroid | Group VII Dispersion |
|---|---|---|---|
| .238 | .000354 | .260 | .000404 |

Pooled W Matrix

| Variable No. | 48 | 55 | 57 |
|---|---|---|---|
| 48 | 5113.92 | 964.272 | -.702948 |
| 55 | 964.272 | 7211.16 | -.010728 |
| 57 | -.702948 | -.010728 | .009995 |

B Matrix

| Variable No. | 48 | 55 | 57 |
|---|---|---|---|
| 48 | 755.612 | 883.732 | .734758 |
| 55 | 883.732 | 1033.64 | .859343 |
| 57 | .734758 | .859343 | .000714 |

Total Correlation Matrix

| Variable No. | 48 | 55 | 57 |
|---|---|---|---|
| 48 | 1.000 | .266 | .004 |
| 55 | .266 | 1.000 | .090 |
| 57 | .004 | .090 | 1.000 |

## LITERATURE CITED

(1) Sharp, A. L., Gibbs, A. E., and Owen, W. J. "Development of a Procedure for Estimating the Effects of Land and Watershed Treatment on Streamflow," U.S. Department of Agriculture Technical Bulletin No. 1352, March 1966.

(2) Fisher, Ronald A. The Use of Multiple Measurements in Taxonomic Problems. Ann. Eugenics, &, Part II, 179-188, 1936.

(3) Rao, C. R. Advanced Statistical Methods in Biometric Research. John Wiley and Sons, New York, 390 pp., 1952.

(4) Kendall, M. G. A Course in Multivariate Analysis. Hafner Publishing Co., New York, 185 pp., 1957.

(5) Anderson, T. W. An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, New York, 374 pp., 1958.

(6) Cooley, W. W., and Lohnes, Paul R. Multivariate Procedures for the Behavioral Sciences. John Wiley and Sons, New York, 211 pp., 1962.

(7) Cosetti, Emilio. "Multiple Discriminant Functions," Northwestern University, Evanston, Illinois, Technical Report No. 11 of ONR Task No. 389-135, Office of Naval Research, Geography Branch, May 1964.

(8) Bryan, J. G. A Method for the Exact Determination of the Characteristic Equation and Latent Vectors of a Matrix with Applications to the Discriminant Function for More than Two Groups. Unpublished Ed. D. dissertation, Graduate School of Education, Harvard University, 290 pp., 1950.

(9) Bryan, J. G. The Generalized Discriminant Functions: Mathematical Foundation and Computational Routine. Harvard Educ. Rev., 21, No. 2, Spring, 90-95, 1951.

(10) Miller, Robert G. "Statistical Prediction by Discriminant Analysis," American Meteorological Society, Meteorological Monographs, Vol. 4, No. 25, October 1962.

(11) Hodges, Joseph L., Jr. Discriminatory Analysis: 1. Survey of Discriminatory Analysis. Report No. 1, USAF School of Aviation Medicine, Randolph Field, Texas, 115 pp., 1950.

(12) Tatsuoka, M. M., and Tiedeman, D. V. "Discriminant Analysis," Review of Educational Research, Vol. XXIV, No. 5, pp. 402-420, 1954.

(13) Hotelling, Harold. "The Generalization of Student's Ratio." Annals of Mathematical Statistics 2: 360-78, August 1931.

(14) Mahalanobis, P. C. "Analysis of Race-Mixture in Bengal." Journal of the Asiatic Society of Bengal, New Series 23: 301-33, 1927.

(15) Welch, B. L. "Note on Discriminant Functions." Biometrika 31: 218-20, July 1939.

(16) Wallace, Noel, and Travers, Robert M. W. "A Psychometric Sociological Study of a Group of Specialty Salesmen.: Annals of Eugenics 8: 266-302, May 1938.

(17) Selover, Robert B. "A Study of the Sophomore Testing Program at the University of Minnesota." Journal of Applied Psychology 26: 296-307, June; 456-67, August; 587-93, October 1942.

(18) Kuder, G. Frederic. Revised Manual for the Kuder Preference Record. Chicago: Science Research Associates, pp. 21-25, 1946.

(19) Baten, William D., and Hatcher, Hazel M. "Distinguishing Method Differences by Use of Discriminant Functions." Journal of Experimental Education 12: 184-86, March 1944.

(20) Harper, A. Edwin, Jr. "Discrimination Between Matched Schizophrenics and Normals by the Wechsler-Bellevue Scale." Journal of Consulting Psychology 14: 351-57, October 1950.

(21) Wherry, Robert J. "Multiple Bi-Serial and Multiple Point Bi-Serial Correlation." Psychometrika 12: 189-95, September 1947.

(22) Rulon, Phillip J. "Distinctions Between Discriminant and Regression Analysis and a Geometric Interpretation of the Discriminant Function." Harvard Educ. Rev., 21, 80-90, Spring 1951.

(23) Rulon, Phillip J. "The Stanine and Separile: A Fable." Personnel Psychology 4: 99-114, Spring 1951.

(24) Tiedeman, David V. "The Utility of the Discriminant Function in Psychological and Guidance Investigations.": Harvard Educ. Rev., 21, 71-80, Spring 1951.

(25) Barnard, Mildred M. "The Secular Variations of Skull Characters in Four Series of Egyptian Skulls." Annals of Eugenics 6: 352-71, December 1935.

(26) Day, Besse B., and Sandomire, Marion M. "Use of the Discriminant Function for More Than Two Groups.": Journal of the American Statistical Association 37: 461-72, December 1942.

(27) Fisher, Ronald A. Statistical Methods for Research Workers. Tenth edition. London, England: Oliver and Boyd, 354 pp., 1946.

(28) Johnson, Palmer O. Statistical Methods in Research. Prentice-Hall, New York, 377 pp., 1949.

(29) Mather, Kenneth. Biometrical Genetics. Methuen and Co., London, England, 158 pp., 1949.

(30) Wilks, Samuel S. "Certain Generalizations in the Analysis of Variance." Biometrika 24: 471-94, November 1932.

(31) Bartlett, M. S. "The Statistical Significance of Canonical Correlations." Biometrika 32: 29-38, 1941.

(32) Bartlett, Maurice S. "Multivariate Analysis." Supplement to the Journal of the Royal Statistical Society 9: 176-90, 1947.

(33) Lohnes, P. R. "Test Space and Discriminant Space Classification Models and Related Significance Tests." Educational and Psychological Measurement 21: 559-574, 1961.

(34) Rulon, P. J., and Brooks, W. D. "On Statistical Tests of Group Differences.: Educational Research Corporation, Cambridge, Mass.

(35) Bartlett, M. S. "Properties of Sufficiency and Statistical Tests." Proceedings of the Royal Society, A 160: 268-282, 1937.

(36) Box, G. E. P. "A Genial Distribution Theory for a Class of Likelihood Criteria.: Biometrika 36: 317-346, 1949.

(37) Wallis, James R. "When Is It Safe to Extend a Prediction Equation? - An Answer Based upon Factor and Discriminant Function Analysis.": Water Resources Research, Vol. 3, No. 2, pp. 375-384, 1967.

(38) Kossack, Carl F. A Handbook of Statistical Classification Techniques, Purdue University, 1964.

(39) duMas, F. M. "The Coefficient of Profile Similarity.": Journal of Clinical Psychology 5: 123-31; April 1949.

(40) Cattell, R. B. "$r_b$ and Other Coefficients of Pattern Similarity." Psychometrika 14: 279-298, 1949.

(41) Cronbach, L. J., and Gleser, G. C. "Assessing Profile Similarity." Psychological Bulletin 50: 456-473, 1953.

(42) Rulon, P. J., Tiedeman, D. V., Longmuir, C. R., and Tatsuoka, M. M. "The Profile Problems: A Methodological Study of the Interpretation of Multiple Test Scores." Educational Research Corporation, Cambridge, Mass.

(43) Brier, Glenn W., and Allen, Roger A. "Verification of Weather Forecasts." Compendium of Meteorology, Ed., T. F. Malone, Boston, American Meteorological Society, pp. 841-848, 1951.

(44) Sanders, Frederick. "The Evaluation of Subjective Probability Forecasts." Scientific Report No. 5, Contract No. AFCRC-TN-58-465, Cambridge, Mass., Massachusetts Institute of Technology, 65 pp., 1958.

(45) Committee for Hydrological Research T.N.O. Recent Trends in Hydrograph Synthesis. Proceedings of Technical Meeting 21, Central Organization for Applied Scientific Research in the Netherlands T.N.O. The Hague, 1966.

(46) deZeeuw, J. W. "Analyse van het afvoer-verloop in gebieden met hoofdzakelijk grondwater afvoer." Mededelingen Landbouwhogeschool (in print).

(47) Makkink, G. F., and van Heemst, H. D. J. "Water Balance and Water Bookkeeping of Regions." Verslagen en mededelingen van de Commissie voor Hydrologisch Onderzoek T.N.O. No. 12, 90-112, 1966.

(48) Kohler, M. A. "Meteorological Aspects of Evaporation Phenomena." General Assembly, Int. Assoc. of Scientific Hydrology, Vol. 3, 421-436, Toronto, September 3-14, 1957.

(49) van Schilfgaarde, J. "Transient Design of Drainage Systems." American Society of Civil Engineering Water Resources Engineering Conference, Mobile, Alabama, March 8-12, 1965.

(50) Kohler, M. A., and Richards, M. M. "Multi-Capacity Basin Accounting for Predicting Runoff from Storm Precipitation." Journal of Geophysical Research 67 (1962), 5187-97.

(51) Linsley, Roy K., Kohler, Max A., Paulhus, Joseph L. H. Applied Hydrology. McGraw-Hill Book Co., Inc., New York, 1949. Appendix A, "Graphical Correlation," 643-653.

(52) Becker, Alfred. "Threshold Considerations and Their General Importance for Hydrologic Systems Investigation," Proceedings of The International Hydrology Symposium, Vol. 1, 1967, 94-102.

(53) Horton, R. E. "Analysis of Runoff Plot Experiments with Varying Infiltration Capacity." Trans American Geophysical Union, Part IV, 693, 1939.

(54) Holtan, H. N. "A Concept for Infiltration Estimates in Watershed Engineering." USDA, Agricultural Research Service, Bulletin 41-51, October 1961.

(55) United States Department of Agriculture, S.C.S. Hydrology, Supplement A, Section 4 of Engineering Handbook. Washington, D.C., 1957.

(56) Kohler, M. A. "Rainfall-Runoff Models." General Assembly of Int. Assoc. of Scientific Hydrology, "Surface Waters," 479-491, Berkeley, August 19-31, 1963.

(57) Wiser, E. H., and van Schilfgaarde, J. "Prediction of Irrigation Water Requirements Using a Water Balance." Proceedings VIth International Congress of Agr. Engr., 121-129, Lausanne, September 21-27, 1964.

(58) Pattison, A. "Synthesis of Rainfall Data," Stanford University, Department of Civil Engineering, Technical Report 40, 1964.

(59) Ramaseshan, S. "A Stochastic Analysis of Rainfall and Runoff Characteristics by Sequential Generation and Simulation," University of Illinois, PhD Thesis, 1964.

(60) Grace, R. A., and Eagleson, P. S. "The Synthesis of Short-Time Increment Rainfall Sequences," Massachusetts Institute of Technology, Department of Civil Engineering, Hydrodynamics Laboratory Report No. 91, 1966.

(61) Slade, J. J., Jr. "An Asymmetric Probability Function," Transactions A.S.C.E., Vol. 101, p. 35, 1936.

(62) Thom, H. C. S. "On the Statistical Analysis of Rainfall Data," Transactions A.G.U., Vol. 21, Part II, p. 490, 1940.

(63) Merriam, C. F. "Long Term Constancy of Rainfall," Paper presented at Annual Meeting of the American Geophysical Union, Section of Hydrology, 1941.

(64) Markovic, R. D. "Probability Functions of Best Fit to Distributions of Annual Precipitation and Runoff," Colorado State University, Hydrology Paper No. 8, August 1965.

(65) Yule, G. U. "On a Method of Studying Time Series Based on Their Internal Correlations," Journal Roy. Stat. Soc., Vol. 108, pp. 208-225, 1945.

(66) Kotz, S., and Neumann, J. "Autocorrelation in Precipitation amounts," Journal Meteorology, Vol. 16, pp. 683-685, Dec. 1959.

(67) Brittan, M. R., et al. "Past and Probably Future Variations in Stream Flow in the Upper Colorado River," Five Parts, University of Colorado, Bureau of Economic Research, October 1961.

(68) Hoel, P. G. Elementary Statistics, Wiley, New York, 1960.

(69) Pattison, A. "Synthesis of Hourly Rainfall Data," Water Resources Research, Vol. 1, pp. 489-498, 1965.

(70) Whitcomb, Margaret, "A Statistical Study of Rainfall Data," M.I.T., Meteor. M.S. Thesis, 1940.

(71) Stidd, C. K. "Cube-Root-Normal Precipitation Distributions," Trans. A.G.U., Vol. 34, No. 1, pp. 31-35, February 1953.

(72) Beals, G. A. "Specification of Daily Precipitation through Synoptic Climatology," M.I.T. Meteor. M.S. Thesis, 1954.

(73) Besson, L. "On the Comparison of Meteorological Data with Chance Results," Monthly Weather Review, Vol. 48, pp. 89-94, 1920, (translated and abridged by E. W. Woolard).

(74) Besson, L. "Sur la Probabilité de la Pluie,: Compte Rendus, Vol. 178, pp. 1743-1745, 1924.

(75) Beer, A., et al. "Sequences of Wet and Dry Months and the Theory of Probability," Quarterly Journal Royal Meteor. Soc., Vol. 72, pp. 74-86, 1946.

(76) Namias, J. "The Annual Course of Month-to-Month Persistence in Climatic Anomalies," Bull. Amer. Met. Soc., Vol. 33, No. 7, pp. 279-285, 1952.

(77) Das, S. C. "The Fitting of Truncated Type III Curves to Daily Rainfall Data," Australian Journal of Physics, Vol. 8, pp. 298-304, 1955.

(78) Kotz, S., and Neumann, J. "On the Distribution of Precipitation Amounts for Periods of Increasing Length," Journ. Geophys. Res., Vol. 68, pp. 3635-3640, 1963.

(79) Brakensiek, D. L. "Fitting a Generalized Log-Normal Distribution to Hydrologic Data,: Trans. A.G.U., Vol. 39, pp. 469 - 473, 1958.

(80) Uttinger, H. "Die Niederschlagsverhaltnisse der Sudschweiz 1901-1940," Ann. Schweiz. Met., Zent-Anst., Zurich, Vol. 82, p. 23, 1945.

(81) Hannan, E. J. "A Test for Singularities in Sydney Rainfall," Australian Journal of Physics, Vol. 8, pp. 289-297, 1955.

(82) Sellers, W. D. "Prediction of Daily Precipitation by Using Statistical Methods," M.I.T. Meteor. M.S. Thesis, 1955.

(83) Williams, C. B. "Sequences of Wet and Dry Days Considered in Relation to the Logarithmic Series," Quart. Journ. Roy. Met. Soc. Vol. 78, No. 335, pp. 91-96, January 1952.

(84) Longley, R. W. "The Length of Wet and Dry Periods," Quart. Journ. Roy. Met. Soc., Vol. 79, p. 520, 1953.

(85) Cooke, D. S. "The Duration of Wet and Dry Spells at Moncton, New Brunswich," Quarterly Journal of the Roy. Met. Soc., Vol. 79, No. 342, pp. 536-538, October 1953.

(86) Gabriel, K. R., and Neumann, J. "A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv," Quart. Journ of the Roy. Met. Soc., London, Vol. 88, pp. 90-95, 1962.

(87) Caskey, J. E., Jr. "A Markov Chain Model for the Probability of Precipitation Occurrence in Intervals of Various Length," Monthly Weather Review, Vol. 91, pp. 298-301, 1963.

(88) Weiss, L. L. "Sequences of Wet and Dry Days Described by a Markov Chain Probability Model," Monthly Weather Review, Vol. 92, pp. 169-176, 1964.

(89) Newnham, E. V. "The Persistence of Wet and Dry Weather," Quart. Journ. Roy. Met. Soc., Vol. 42, No. 179, pp. 153-162, July 1916.

(90) Jorgensen, D. L. "Persistency of Rain and No-Rain Periods During the Winter of San Francisco," Monthly Weather Review, Vol. 77, pp. 303-307, 1949.

(91) Wiser, E. H. "Modified Markov Probability Models of Sequences of Precipitation Events," Monthly Weather Review, Vol. 93, pp. 511-516, 1965.

(92) Feyerherm, A. M., and Bark, L. D. "Statistical Methods for Persistent Precipitation Patterns," Journ. Appl. Met., Vol. 4, pp. 320-328, 1965.

(93) Green, J. R. "A Model for Rainfall Occurrence," Journ. of the Roy. Stat. Soc., Series B, Vol. 26, pp. 345-353, 1964.

(94) Enger, I. "Some Attempts at Predicting a Meteorological Time Series from Its Past History," M.I.T. Meteor. M.S. Thesis, 1957.

(95) Hammerle, J. F. "Linear Prediction of Discrete Stationary Time Series," M.I.T. Math. M.S. Thesis, 1951.

(96) U.S. Department of Agriculture. The Agriculture, Soils, Geology, and Topography of the Blacklands Experimental Watershed, Waco, Texas.

(97) Hartman, M. A. "Determining Rainfall-Runoff-Retention Relationships." Texas Agricultural Experiment Station, MP 404, January 1960.

(98) Hartman, M. A., et al. "Precipitation-Soil Moisture Relations for the Blacklands of Texas." Hydrology Workshop of the Soil Conservation Service, Forest Service, and Agricultural Research Service, New Orleans, Louisiana, October 24-27, 1960.

(99) U. S. Geological Survey, Water-Loss Investigations: Lake Hefner Studies, Technical Report. Professional Paper 269, 1954.

(100) Texas Agricultural Experiment Station, Water Evaporation Studies in Texas, Bulletin 787, p. 29, November 1954.

(101) Corps of Engineers, U. S. Army, Simulation of Monthly Runoff, Technical Bulletin No. 1, Hydrologic Engineering Center, U.S. Army Engineer District, p. II-3, November 1964.

(102) Fiering, Myron B. Streamflow Synthesis. Harvard University Press, Cambridge, Mass., 1967.

(103). MacLaren, M. D., and Marsaglia, G. Uniform Random Number Generators. Journal Association for Computing Machinery, Vol. 12, pp. 83-89, 1965.

(104) Van Gelder, A. Some New Results in Pseudo-Random Number Generation. Journal Association for Computing Machinery, Vol. 14, No. 1, pp. 785-793, October 1967.

(105) Lindgren, B. W., and McElrath, G. W. Introduction to Probability and Statistics. Macmillan Company, New York, 1959.

(106) Hoel, Paul G. Introduction to Mathematical Statistics, Third edition, John Wiley and Sons, New York, 1962.

# OTHER REFERENCES

## General

(1) Allard, J. L., Dobell, A. R., and Hull, T. E. "Mixed Congruential Random Number Generators for Decimal Machines." Journal Association for Computing Machinery, Vol. 10, pp. 131-141, 1963.

(2) Brown, George W. "Discriminant Functions." Am. Math. Stat. 18, pp. 514, 528, December 1947.

(3) Cox, D. R., and Miller, H. D. The Theory of Stochastic Processes, John Wiley and Sons, New York, 1965.

(4) Crow, Edwin L., Davis, Frances A., and Marfield, Margaret W. Statistics Manual, Dover Publications, Inc., New York, 1960.

(5) Fraser, D. A. S. Nonparametric Methods in Statistics, Fifth Printing, John Wiley and Sons, New York, 1966.

(6) Hoel, Paul G. Introduction to Mathematical Statistics, Third Edition, John Wiley and Sons, 1962.

(7) Parzen, Emanuel. Stochastic Processes, Holden-Dorf, Inc., San Francisco, 1962.

(8) Quenouille, M. H. Associated Measurements, Butterworths Scientific Publications, London, 1952.

(9) Roy, S. N. Some Aspects of Multivariate Analysis, Calcutta; Indian Statistical Institute, John Wiley and Sons, New York, 1957.

(10) Seal, Hilary L. Multivariate Statistical Analysis for Biologists, John Wiley and Sons, Inc., New York, 1964.

(11) Inland Waters Branch, Department of Energy, Mines and Resources, Statistical Methods in Hydrology. Proceedings of Hydrology Symposium No. 5, Queen's Printer and Controller of Stationery, Ottawa, Canada, 1967.

## Factor Analysis

(12) Anderson, T. W. An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, New York, 1958.

(13) Cooley, William M., and Lohnes, Paul R. Multivariate Procedures for The Behavioral Sciences, John Wiley and Sons, New York, 1965.

(14) Harman, Harry H. Modern Factor Analysis, The University of Chicago Press, Chicago, 1960.

(15) Kendall, M. G. A Course in Multivariate Analysis, Griffins Statistical Monographs and Courses, Hafner Publishing Co., New York, 1957.

## Recent Use of Multiple Discriminant Analysis

### Biological Abstracts - Discriminant Analysis and Function

1966

(1) The Heritage of Hypertension and Coronary Disease: Discriminant Function Analysis of the Characteristics of Healthy Young Adults, by Caroline Bedell Thomas and Donald Clare Ross, Amer. J. Med. Sci. 248(5): 505-513. Illus. 1964.

(2) Precursors of Hypertension and Coronary Disease Among Healthy Medical Students: Discriminant Function Analysis II, Using Parental History as the Criterion, by Caroline Bedell Thomas, Donald Clare Ross, and Carolyn Q. Higinbothom. Bull. Johns Hopkins Hosp. 115(3), 245-264. Illus. 1964.

(3) Complex Behavioral Indices Weighted by Linear Discriminant Functions for the Prediction of Cerebral Damage, by Lawrence Wheeler. Percept. Mot. Skills, 19(3): 907-923. Illus. 1964.

(4) Estimation of Error Rates in Discriminant Analysis, by Peter Anthony Lachenbruck. Diss. Abst. 26(1): 325. Abstract only, 1965.

(5) Discriminant Function Analysis as an Aid to the Diagnosis of Flax Byscinosis, by J. D. Merrett. Brit. J. Ind. Med. 23(1): 58-61, 1966.

(6) Computing a Discriminant Function from Within-Sample Dispersions, by M. J. R. Healy. Biometrics 21(4): 1011-1012, 1965.

1965

(7) Use of a Discriminant Function to Classify North American and Asian Pink Salmon, Orcorhynchus Gorbuscha (Walbaum), collected 1959, by Roger E. Pearson, Int. in Pacific Fish Comm. Bull. 14, 67-90. Illus. 1964.

(8) A Discriminant Function Analysis of Articulation in Cleft-Palate Prothesis [Man], Charles R. Elliott, Asha J. Amer. Speech Hearing Assoc. 6(10): 394. Abstract only, 1964.

(9) Sex Determination by Discriminant Function Analysis of the Mandible, by Eugene Giles, Amer. J. Phys. Anthropol. 22(2): 129-135. Illus. 1964.

(10) Multiple Discriminant Analysis of Plasma Amino Acid Patterns, by Frank L. Siegel, Proc. Nat. Acad. Sci. U.S.A. 51(5): 866-871. Illus. 1964.

(11) The Use of Fisher's Discriminant Function in the Taxonomy of Small Algae (Oocystis A. Braun), by Pavel Javornicky and Helena Rehakova, Preslia (praha) 36(2): 105-113. Illus. 1964. Czeck.

(12) Discriminant Function in Wheat Hybrid, by V. S. Bhide, Indian Agric. 7(1/2): 76-78, 1963.

(13) Precursors of Hypertension and Coronary Disease Among Healthy Medical Students: Discriminant Function Analysis I, Using Smoking Habits as the Criterion, by C. B. Thomas, D. C. Ross, and C. Q. Hinginbothem, Bull. Johns Hopkins Hosp. 115(2): 174-194. Illus. 1964.

1964

(14) Myocardial Infarction Prognosis by Discriminant Analysis, by William L. Hughes, John M. Kalbfleisch, Edward N. Bramdt, Jr., and J. Paul Costiloe, Arch. Internal Med. 111(3): 338-345. Illus. 1963.

(15) Use of a Discriminant Function in the Morphological Separation of Asian and North American Races of Pink Salmon, Oncorhynchus Gorbuscha (Walbaum), by Murray H. Amos, Raymond E. Amos, and Roger E. Pearson, Int. N. Pacific Fish Comm. Bull. 11, 73-100. Illus. Maps. 1963.

(16) Parents of Schizophrenic Children Compared with the Parents of Non-Psychotic Emotionally Disturbed and Well Children: A Discriminant Function Analysis, by Bernard Cooper, Dissert. Abst. 24(4): 1694-1695. Abstract only, 1963.

(17) Potential Water Deficite as a Climatic Discriminant, by F. H. W. Green, The Water Relations of Plants by O. J. Rulter and F. W. Whitehead, John Wiley, 1963.

1963

(18) On the Use of Discriminant Functions in Taxonomy, by Alexander A. Lubischew, Biometrics 18(4): 455-477. Illus. 1962.

(19) Discriminant Functions in Psychobiometric Investigation, by Felipe Montemayor and Maria Teresa Jain, Am. Inst. Nac. Antrop. e Hist. 11(40): 219-242. Illus. 1957/1958.

(20) Linear Discriminant Analysis in Perinatal Mortality, by Bernard G. Greenberg and H. Bradley Wells, Amer. Jour. Publ. Health 53(4): 594-602. Illus. 1963.

(21) Sex Determination by Discriminant Analysis of Crania, by Eugene Giles and Orville Elliott, Amer. Jour. Phys. Anthropol. 21(1): 53-68. Illus. 1963.

1962

(22) New Methods of Quantitative Representations on the Structure of Plan Communities I. Application of Discriminating Equation of Polyo-Eggenberger Distribution, by Kayama Ryosei, Japanese Jour. Ecol. 11(1): 4-10. Illus. 1961. Eng. Seminar.

(23) The Psychodiagnostic Efficiency of WAIS and Rorschach Scores: A Discriminant Function Study, by Robert Lee Geiser, Diss. Abstr. 22(3): 915. 1961. Abstract only.

(24) Discriminant Analysis in Biometrical Genetics, S. K. Jain, Nature 191(4796): 1420. 1961.

(25) Sexing Indian Crania by Discriminant Analysis. In Proc. of 13th Ann. Meeting Amer. Assoc. of Physical Anthropologists, Columbus, Ohio, May 1961 by Eugene Giles and Orville Elliott, Amer. Jour. Phys. Anthropol. 20(1): 64. 1962 abstract.

1961

(26) The Maximum Separation of Students into Two Programs of Course Work in a College of Agriculture by Discriminant Analysis Involving Certain Selected Measurements, by Gilf Soquiguit, Philippine Agri. 44(4): 149-196. Illus. 1960.

(27) Value of Drumsticks and Other Nuclear Appendices in the Determination of Sex. Discriminatory Analysis Based on Findings of 804 Normal Subjects, by J. Procopio-Valle, Walker A. Chagas, and J. N. Manceaw, Jour. Clin. Endocrinal. and Metab. 216(8): 9650975. Illus. 1961.

1960

(28) Correlation Studies and the Application of Discriminant Function in Aestivum Wheats for Varietal Selection Under Rainfed Conditions by S. M. Sikka and K. B. L. Jain, Indian Jour. Genetics and Plant Breeding 18(2): 178-186. 1958.

1954

(29) Statistics of Discrimination in Anthropology, by J. Bronowski and W. M. Long, Amer. Jour. Phys. Anthropol. 10(4): 385-394, 1952.

(30) An Application of the Linear Discriminant Function to Insect Taxonomy, by R. S. Bigelow and C. Reimer, Canadian Ent. 86(2): 69-73, 1954.

1950
(31) A Further Note of Discriminatory Analysis, by M. H. Quenouille, Am. Engenics. 15(1): 11-14. 1949.

1949
(32) An Application of Discriminant Function for Selection in Durum Wheats, By K. M. Simlote, Indian Jour. Agric. Sci. 17(5): 269-279, 1947.

1948
(33) Discriminant Functions, By G. W. Brown, Am. Math. Stat. 18(4): 514-528, 1947.

1947
(34) Some Notes on Discrimination, By L. S. Penrose, Am. Eugenics, 13: 228-237, 1947.

(35) Some Examples of Discrimination, by Cedric A. B. Smith, Am. Eugenics, 13: 272-282, 1947.

1946
(36) The Use of Discriminant Functions in Comparing Juges Scores Concerning Potatoes, by W. D. Baten, Jour. Amer. Statistic Assoc. 40(230): 223-228, 1945.

Psychological Abstracts

1965
(1) Complex Behavioral Indices Weighted by Linear Discriminant Functions for the Prediction of Cerebral Damage, by Lawrence Wheeler, Perceptual and Motor Skills, 19(3): 907-923, 1964. (Also listed in Biological Abstracts).

(2) Multiple Discriminant Prediction of Delinquency and School Drop-outs, by Francis J. Kelly, Donald J. Veldman, and Carson McGuire, Educational and Psychological Measurement, 24(3): 535-544, 1964.

1964
(3) A Method for Detecting Subgroups in a Population and Specifying Their Membership, by J. A. Gengerelli, J. Psychol., 55(2): 457-468, 1963.

(4) An Application of Discriminant Functions to the Problem of Predicting Brain Damage Using Behavioral Variables, by L. Wheeler, C. J. Burke, and R. M. Reitan, Percept. Mot. Skills, 16(2): 417-440, 1963.

1963
(5) The Discriminant Function Analysis: Its Technique and Use in Psychology and Psychiatry, by N. Sundararaj, Pratibha 2(1): 66-70, 1959.

(6) Some Remarks on Failure to Meet Assumptions in Discriminant Analyses, by Richard S. Melton, Psychometrika 28(1): 49-53, 1963.

1962
(7) Subjective Discrimination as a Statistical Method, by M. Stone, Brit. J. Statist. Psychol. 14, 25-28, 1961.

1961
(8) The Discriminant Function Analysis: Its Technique and Use in Psychology and Psychiatry, by N. Sundararaj, J. All-India Inst. Mental Health, 2(1): 66-70, 1959.

(9) An Examination of the Use of Linear and Nonlinear Discriminant Functions in the Classification of College Students into Academic Groups, by Walter R. Stellwagen, Dissertation Abstracts 20, 4435, Abstract only, 1960.

(10) Multiple Discriminant Analysis Applied to "Ways to Live" Ratings from Six Cultural Groups, by L. V. Jones and R. D. Bock, Sociometry 23, 162-176, 1960.

(11) The Use of Discriminant Analysis in the Prediction of Scholastic Performance, by Vincent F. Calia, Personnel Guid. J. 39, 184-185, 1960.

(12) Notes on the Use of Discriminating Functions, By A. Lubin, Bull. Cent. Etud. Rech. Psychotech. 9, 63-70, 1960.

(13) The Discriminant Analysis Technique in Assigning Freshmen to College Chemistry Courses, by E. M. Rusten and A. C. F. Gilbert, J. Psychol. Stud., 11, 253-255, 1960.

1957
(14) The Comparability of the Simple Discriminant Function and Multiple Regression Techniques, by William B. Michael and Norman C. Perry, J. Exp. Educ. 24, 299-301, 1956.

1956
(15) An Application of Fisher's Discriminant Function in the Classification of Students, by Stanley J. Ahmann, J. Educ. Psychol. 46, 184-188, 1955.

1954
(16) The Predictive Use of the Linear Discriminant Function in Naval Aviation Cadet Selection, by Robert F. Lockman, U.S. Naval Sch. Aviat. Med. Res. Rep. 1953, Rep. No. N.M. 001057.16.02.11p.

1953
(17) The Use of Discriminant Functions in Individual Diagnosis in Psychology, by P. Pichot and J. Perse, Rev. Psychol. Appl. 2, 19-34, 1952.

1952
(18) Inter-Tests and Scatter Comparisons in Psychopathology; Methods and Perspectives, by Pierre Pichot, Arech. Psicol. Neurol. Psichiat. 12, 304-310, 1951.

(19) The Goodness of Fit of a Single Hypothetical Discriminant Function in the Case of Several Groups, by M. S. Bartlett, Am. Eugen., Comb. 16, 199-214, 1951.

1951
(20) A Comparison of Two Procedures for Calculating Discriminant Function Coefficients, by John Schmid Jr., Psy-chometrika, 15, 431-434, 1950.

(21) The Variance Error of the $P_{50}$ - Discriminant, by Gilbert L. Betts, Psychometrika, 15, 435-439, 1950.

(22) Basic Principles for Construction and Application of Discrimination by George W. Brown, J. Clin. Psychol. 6, 58-61, 1950.

1948
(23) A Statistical Criterion to Determine the Group to which an Individual Belongs, by Rodhakrishna C. Rao, Nature, London, 1947, 160, 835-836.

(24) Some Notes on Discrimination, by L. S. Penrose, Am. Eugen. Comb. 13, 228-237, 1947. (See Biological Abstracts also).

(25) Some Examples of Discrimination, by Cedric A. B. Smith, Ann. Eugen. Comb. 13, 272-282, 1947. (See Biological Abstracts also).

# VITA

## General

Donn Gene DeCoursey - 120 Ruskin Place, Chickasha, Oklahoma 73018

Born - Auburn, Indiana Cotober 21, 1934

Parents - Mr. E. M. DeCoursey, Mechanical Engineer (Deceased)
Mrs. Bessie DeCoursey, Elementary School Teacher
R. R. #3, Auburn, Indiana 46706

Wife - Mrs. Shirley Ann DeCoursey, Homemaker
B.S. Home Economics, Purdue University 1957

Children - Kenneth Allen age 11
Louann Elizabeth age 3

Social Interests - First Presbyterian Church, Chickasha, Oklahoma
Deacon
Boy Scout Troop 301, Assistant Scout Master
Little League, Assistant Coach
Toastmasters Club, Chickasha, Oklahoma

## Education

High School - McIntosh, Auburn, Indiana Graduated 1952

Bachelor of Science - Purdue University - Agricultural Engineering, June 1957
Major - Agricultural Engineering, emphasis in soil and water
Minor - Physics, Mathematics

Master of Science - Purdue University - Agricultural Engineering
June 1958
Major - Hydrology, Hydraulics

| | | |
|---|---|---|
| Drainage, Flood Control, Erosion | 3 | sem. hrs. |
| Hydrology | 3 | " " |
| Applied Hydraulics | 3 | " " |
| Design of Hydraulic Structures | 3 | " " |

Minors - Soil Mechanics

| | | |
|---|---|---|
| Soil Mechanics | 3 | " " |
| Ground Water and Seepage | 3 | " " |
| Design of Earth Structures | 3 | " " |

Statistics
Advanced Statistical Methods 3 sem. hrs.
Correlation - Regression 3 " "
Thesis - Accuracy of Runoff Measuring Systems

Doctor of Philosophy - Georgia Institute of Technology - Civil Engineering, anticipated June 1970
Major - Hydrology
Hydrologic Models 3 qtr. hrs.
Watershed Analysis 3 " "
Urban Hydrology 3 " "
Meteorology 3 " "
Hydrometeorology - Flood Synthesis 3 " "
Hydraulics
Intermediate Fluid Mechanics 3 " "
Steady Flow in Open Channels I 3 " "
Steady Flow in Open Channels II 3 " "
Flow in Enclosed Conduits 3 " "
Mechanics of Flow in Porous Media 3 " "
Sediment and Sediment Transport 3 " "
Hydrodynamics audit
Water Resources
Technology in Water Resources Development 3 qtr. hrs.
Economics in Water Resources Development 3 " "
Seminar in Water Resources Engineering 3 " "
Minor - Statistics and Mathematics
Advanced Engineering Mathematics 3 " "
Multivariate Computer Methods 3 " "
Methods of Operations Research 5 " "
Research - Use of Multiple Discriminant Analysis to Evaluate the Effects of Land Use on the Simulated Yield of a Watershed

## Professional Experience

Feb. 1958 - April 1960, Engineer, Indiana Flood Control and Water Resources Commission, Indianapolis, Indiana. Principal area of work; hydrologic analyses.

April 1960 - May 1960, Hydraulic Engineer, E. R. Hamilton Associates, Inc., Indianapolis, Indiana. Principal area of work; design of a sewage treatment plant.

June 1960 - June 1961, Hydraulic Engineer, Indiana Flood Control and Water Resources Commission, Indianapolis, Indiana. Principal area of work; hydrologic analyses.

June 1961 - Present, Hydraulic Engineer (Research), USDA, ARS, Soil and Water Conservation Research Division, Chickasha, Oklahoma. Principal areas of work; hydrologic analyses, watershed modeling, multivariate statistical analyses, supervision of data collection and analysis.

## Major Publications

DeCoursey, Donn G. Water yield computations. Transactions, ASAE, Vol. 8, No. 3, pp. 367-370, 1963.

DeCoursey, Donn G. A runoff hydrograph equation. USDA, ARS 41-116, February 1966.

Schoof, Russell R., and DeCoursey, Donn G. Conveyance of irrigation water in a natural channel. Proceedings, Second Annual American Water Resources Conference, Chicago, Illinois, November 20-22, 1966.

DeCoursey, Donn G. An application of computer technology to hydrologic model building. IASH Symposium on the Use of Analog and Digital Computers in Hydrology, Tucson, Arizona, Pub. No. 80, AIHS Vol. 1, December 1968.

Shockey, Windell R., and DeCoursey, Donn G. Point sampling of land use in the Washita Basin. USDA, ARS 41-149, April 1969.

Blanchard, Bruce J., and DeCoursey, Donn G. A technique for measuring rapidly changing streamflow. USDA, ARS 41-145, November 1968.

DeCoursey, Donn G., and Snyder, Willard M. Computer-oriented method of optimizing model parameters. Journal of Hydrology 9 (1969) pp. 34-56; North Holland Publishing Co., Amsterdam, the Netherlands.

DeCoursey, Donn G., and Blanchard, Bruce J. Flow analysis of a large triangular weir. Pending publication in the Journal of the Hydraulics Division, American Society of Civil Engineers.

DeCoursey, Donn G., and Seely, Edward H. Indirect determination of synthetic runoff. XIIIth Congress of the International Association for Hydraulic Research, Proceedings Vol. 1 (Subject A), Kyoto, Japan, August 31 - Sept. 5, 1969.

Blanchard, Bruce J., and DeCoursey, Donn G. A design for low flow controls, pending publication in American Water Resources Association.

## Presentations

Water yield computations. Presented at National meeting of the American Society of Agricultural Engineers, New Orleans, Louisiana, December 1964.

Application of a runoff hydrograph equation. Presented at Fifth Western National Meeting, American Geophysical Union, Dallas, Texas, September 1965.

An application of computer technology to hydrologic model building. Presented at the International Association of Scientific Hydrology, Symposium on Analog and Digital Computers in Hydrology, Tucson, Arizona, December 1968.

Effects of flood retention structures on the flow regime of the Washita River. Presented at the International Arid Lands Conference, Tucson, Arizona, June 1969.

Flow analysis of a large triangular weir. Presented at the ASCE Water Resources Symposium, Memphis, Tennessee, January 1970.

A critique on stochastic modeling. Presented at an ARS-SCS National Workshop on Watershed Modeling, Tucson, Arizona, March 15, 1970.

## Society Memberships

American Geophysical Union (Hydrology Section)

American Society of Civil Engineers

American Society of Agricultural Engineers

International Association for Hydraulic Research

Registered Professional Engineer, State of Indiana No. 9403

Society of The Sigma Xi

## Miscellaneous Information

Member of: Surface Water Committee, Section of Hydrology, American Geophysical Union

Parametric Hydrology Committee of the International Association of Scientific Hydrology, International Union of Geodesy and Geophysics.

Attended Short Course on Linear Theory of Hydrologic Systems conducted by James C. I. Dooge, sponsored by USDA Hydrograph Laboratory and the University of Maryland, College Park, Maryland, August 21 - September 1, 1967.